The Dissertation Committee for Hsin-Chan Huang
certifies that this is the approved version of the following dissertation:

Stockpiling and Resource Allocation for Influenza Preparedness and Manufacturing Assembly

Committee:

David P. Morton, Supervisor

Erhan Kutanoglu, Co-Supervisor

John Hasenbein

Jonathan F. Bard

Lauren A. Meyers

Stockpiling and Resource Allocation for Influenza Preparedness and Manufacturing Assembly

by

Hsin-Chan Huang, B.S.E., M.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2014

Dedicated to my family.

Acknowledgments

I wish to thank the multitudes of people who helped me. First, I would like to thank Dr. David Morton for being my advisor. I have learned a lot and am still learning from you. I would like to thank Dr. Erhan Kutanoglu for being my co-advisor. I have enjoyed meeting with you and learning from you. I would like to thank Dr. John Hasenbein for advising me after Dr. Kutanoglu went to Turkey. Every time when we have a meeting, I learn things from you. I would like to thank Dr. Jonathan Bard and Dr. Lauren Meyers for serving on my committee. Your helpful suggestions made this dissertation more comprehensive. I would like to thank Dr. Dragan Djurdjanovic for being a bystander in my defense. Your helpful suggestions glued the chapters more tightly. Chapters 2 and 3 involve work that was part of two larger collaborative projects for the Texas Department of State Health Services (DSHS). I would like to thank the entire team that worked on these projects, especially Dr. Morton, Dr. Meyers, Dr. Ozgur Araz, Gregory Johnson, Bismark Singh, and Bruce Clements and his team in DSHS. Without everyone's dedicated work, these projects could not have been accomplished so nicely. Chapter 4 is based on a project for an engine assembly company. I would like to thank everyone who was involved in the project, especially Dr. Kutanoglu, James Oliphant, and John Tackett and his team in the company. Without everyone's help, the performance of the assembly lines could not have been improved so much. I would like to thank my friends and colleagues at UT Austin. I will have many great memories of Austin and the ORIE program. Finally, I must thank my wife, my mom, Uncle Lai, Aunty Shiiu, and my family. Your unconditional support helped me go through every step of my graduate study.

Stockpiling and Resource Allocation for Influenza Preparedness and Manufacturing Assembly

Publication No. _____

Hsin-Chan Huang, Ph.D.
The University of Texas at Austin, 2014

Supervisors: David P. Morton
Erhan Kutanoglu

Stockpiling resources is a pervasive way to handle demand uncertainty and future demand surges. However, stockpiling is subject to costs, including warehousing costs, inventory holding costs, and wastage of expired resources. Hence, how to stockpile in an economically efficient manner is an important topic to study. Furthermore, if the inventoried supply is insufficient for a surge in demand, how to best allocate available resources becomes a natural question to ask. In this dissertation, we consider three applications of stockpiling and resource allocation: (i) we stockpile ventilators both centrally and regionally for an influenza pandemic; (ii) we allocate limited vaccine doses of various types to target populations for an influenza pandemic; and, (iii) we investigate inventory needs for low cost, high usage (class C) parts in an engine assembly plant.

First, we describe and analyze a model for estimating the number of ventilators that the Texas Department of State Health Services (DSHS), and eight health

service regions in Texas, should stockpile for an influenza pandemic. Using a probability distribution governing peak-week demand for ventilators across the eight health service regions, an optimization model allows investigation of the tradeoff between the cost of the total stockpile and the expected shortfall of ventilators under mild, moderate, and severe pandemic scenarios. Our analysis yields the surprising result that there is little benefit to DSHS holding a significant stockpile, even when those centrally held ventilators can be dispatched to regions *after* observing the peak-week demand realization. Three factors contribute to this result: positively correlated regional demands, a relatively low coefficient of variation, and wastage of the central stockpile once it is dispatched to the regions.

Second, we formulate an optimization model for allocating various types of vaccines to multiple priority groups in 254 counties in the state of Texas that DSHS can use to distribute its vaccines for an influenza pandemic. For reaching the public, vaccines are allocated to the state's Registered Providers (RPs), Local Health Departments (LHDs), and Health Service Regions (HSRs). The first two allocations are driven by requests from RPs and LHDs while HSR allocation is at DSHS's discretion. The optimization model aims to achieve proportionally fair coverage of priority groups across the 254 counties, as informed by user-specified weights on those priority groups, using the HSR doses. With proportional fairness as our primary goal, the optimal allocation also counts policy simplicity and regional equity. Sensitivity analysis on the portion of the state's vaccines reserved for HSRs shows that a small portion can effectively shrink the gap of vaccination coverage between urban and rural counties.

Finally, we derive short-cut formulae for estimating the extra inventory needed

for managing class C parts in units of bins that an engine assembly plant can use to achieve a desired fill rate at workstations. The plant orders a class C part from its supplier based on the part's aggregated next-day demand across all workstations. After receiving the part, the plant first stores the supply in the warehouse and delivers the part to workstations in bins whenever the line-side inventory at a workstation is empty. We study four cases of various information availability in the order quantity calculation and derive associated formulae for estimating the extra inventory needed due to demand aggregation and bin delivery. We demonstrate the performance of our short-cut formulae, showing the tradeoff between extra inventory needed and the associated risk of not satisfying all workstation requests. Our sensitivity analysis shows that workstation demand variation and bin size have little or no influence on the performance of our short-cut formulae.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Stockpiling resources is a common way to handle demand uncertainty and future surges in demand, especially in influenza preparedness and manufacturing assembly. When an influenza pandemic occurs, a surge in demand for medical resources, e.g., vaccines, antivirals, and ventilators, is required for prophylactic treatments and for treating those who have contracted the virus, in part in an effort to control the pandemic. Without thoughtful preparedness and effective countermeasures, an influenza pandemic can cause great damage; e.g., the 1918 pandemic caused an estimated 30-50 million deaths globally [66]. Due to the difficulty of predicting the timing and severity of next pandemic, as well as the time needed for manufacturing these medical resources, stockpiling has become one of the main components in influenza preparedness [54]. In manufacturing assembly, having all components ready at workstations is essential for workflow continuity, especially in a mixed-model assembly line. Lacking a component, either a large expensive part or a small cheap screw, at a workstation can interrupt the workflow, delay jobs at other workstations, and even cause the whole assembly line to shut down. Several factors, including fluctuating demand, large-quantity discounts, and uncertain supply, have led manufacturers to stockpile inventory [4].

However, stockpiling comes with costs, including warehousing costs, inventory holding costs, and wastage of expired resources. Hence, how to stockpile in an eco-

nomic manner has become an important topic to study. The notion of risk pooling is pervasive in protecting against financial losses and in controlling costs to satisfy demand in supply-chain management. Through aggregating individual stochastic demands, we can often dramatically decrease the amount of resources that we must stockpile to limit the level of risk. In practice, risk pooling is often implemented by centralizing inventory, albeit with an associated cost of dispatching centralized inventory to places where it is needed [49]. For influenza preparedness, the Strategic National Stockpile (SNS) program of the U.S. Centers for Disease Control and Prevention (CDC) [62] holds large quantities of medical resources stored in strategically located warehouses for public health emergencies, such as an influenza outbreak. Once an emergency is declared, the SNS resources are delivered to states in need within hours or days, depending on the type of resource. In a manufacturing assembly plant, it is not rare that a component, e.g., bolts and nuts, may be used at several workstations with various demands. For inventory reduction, the plant may store such common components in a centralized location after receiving the supply and then deliver them to workstations upon receiving workstation requests.

Furthermore, if the inventoried supply is insufficient for a surge in demand, how to best allocate available resources becomes a natural question to ask. When allocating scarce life-or-death resources, fairness, and other ethical considerations, are often often used as criteria, in addition to maximizing use of the resources. In allocating limited medical resources for an influenza pandemic, federal authorities often seek to provide equal access across geographic areas. For example, the CDC allocated H1N1 vaccines to states of the U.S. in proportion to their populations during the 2009 H1N1 pandemic [53]. On the other hand, local healthcare providers may use a patient-specific metric to decide who receives a limited critical medical resource. For

example, the Texas Department of State Health Services [58] recommends healthcare facilities using the sequential organ failure assessment (SOFA) score to assign limited mechanical ventilators to patients during an influenza pandemic. In dispatching centralized inventory to workstations, an manufacturing assembly plant often follows a first-come-first-served policy. That is, the plant dispatches inventory to workstations according to the order of workstation requests. Nevertheless, the plant may move components from workstation to workstation after dispatching if needed.

Maximizing system performance given limited resources is a pervasive theme in operations research, as is the dual problem of assessing the minimum required resources to obtain a requisite level of system performance. In this dissertation we consider three problems of stockpiling and resource allocation: (i) we stockpile ventilators both centrally and regionally for an influenza pandemic; (ii) we allocate limited vaccine doses of various types to target populations for an influenza pandemic; and, (iii) we investigate the inventory needed for low cost, high usage (class C) parts in an engine assembly plant.

First, we formulate a two-stage stochastic program for stockpiling ventilators centrally and regionally for an influenza pandemic in the state of Texas. The Texas Department of State Health Services (DSHS) manages a centrally held stockpile. DSHS partitions Texas into eight heath service regions (HSRs), and hospitals in these HSRs hold inventories of ventilators that we aggregate in our model to eight regionally held stockpiles. These centrally and regionally held inventories represent first stage variables in our stochastic program because they must be selected prior to knowing regional demand for patients requiring mechanical ventilation due to pandemic influenza. We make use of a Bayesian dynamic linear model (DLM) with weekly

time increments developed in [23]. Ventilators are reusable commodities, and hence the peak-week's demand, as opposed to cumulative demand, drives proper inventory levels. We model the peak-weak demand in each HSR using a multivariate normal distribution whose parameters are informed by the DLM and by a spatial correlation analysis. Given samples drawn from this multivariate demand distribution, we construct a sample average approximation of the two-stage stochastic program, which allows us to analyze the tradeoff between total stockpile and expected shortfall of ventilators under three pandemic scenarios: mild, moderate, and severe. We perform sensitivity analysis with respect to our model's input parameters to obtain insights as to how they affect stockpiling strategies.

Second, we formulate an optimization model for allocating various types of vaccines to multiple priority groups in 254 counties in the state of Texas. For serving the public, vaccines are allocated to the state's Registered Providers (RPs), Local Health Departments (LHDs), and Health Service Regions (HSRs). The first two allocations are driven by requests from RPs and LHDs while the HSR allocation is at DSHS's discretion. In 2009, these discretionary HSR doses were largely used to boost the coverage of rural counties where an insufficient number of doses were distributed to RPs. The novel optimization model that we construct takes as input all doses allocated to date to RPs, LHDs, and HSRs, and recommends as output the allocation of available discretionary HSR doses targeted to priority groups by vaccine type, all at the geographic resolution of counties. The optimization model aims to achieve proportionally fair coverage of priority groups across the 254 counties, as informed by user-specified weights on those priority groups, using the HSR doses. With proportional fairness as the primary goal, the optimal allocation also accounts for policy simplicity and regional equity. In a retrospective analysis of the 2009 H1N1

pandemic, we simulate the potential use of the 7% of total doses reserved for allocation to HSRs at DSHS's discretion. We also perform sensitivity analysis on the number of vaccines reserved for HSRs to see how it affects the coverage rates between rural counties and urban counties.

Third, we derive short-cut formulae for estimating the extra inventory needed for managing class C parts in units of bins that an engine assembly plant with multiple lines can use to achieve a desired fill rate at workstations. A class C part is usually a small, relatively inexpensive part and hence purchased in bulk and assembled into engines across multiple workstations in varying quantities. The plant orders each part in bins of different quantities from its supplier based on the part's aggregated next-day demand across all workstations. The parts are stored in the plant warehouse before they are requested from workstations. A workstation's line-side inventory is replenished in bins from the plant warehouse whenever the inventory becomes empty. It is not uncommon to have an excessive amount of inventory at one station while another station suffers from part unavailability because all the plant warehouse inventory has been allocated to other stations. The plant implements a concept of a minimum inventory level for each part, which tries to reduce the effect of demand aggregation and bin delivery by ensuring at least a certain amount of inventory remaining in the plant at the end of every day. We study four cases of various information availability in the order quantity calculation and derive the associated formulae for estimating the extra inventory needed to control the risk of not satisfying demands at each workstation due to demand aggregation and bin delivery. We use numerical examples to demonstrate the performance of our short-cut formulae and show the tradeoff between the minimum inventory required and the associated risk. We also perform sensitivity analysis on workstation demand variation and bin size to

see if theses two factors affect the performance of our short-cut formulae.

The reminder of this dissertation is organized as follows. In Chapter 2, we describe how to stockpile ventilators both centrally and regionally for an influenza pandemic, exploring the tradeoff between total stockpile and expected shortfall of ventilators. In Chapter 3, we discuss how to allocate available vaccine doses of various types to target populations, seeking proportionally fair coverage across all 254 counties in the state of Texas. In Chapter 4, we detail how we derive short-cut formulae for estimating the extra inventory needed for class C parts due to workstation demand aggregation and bin delivery. Finally, we conclude in Chapter 5. Chapters 2, 3, and 4 are self contained with a detailed introduction, related literature review, modeling framework, analysis, and discussion, so that a reader can read any of these chapters alone.

Chapter 2

Stockpiling Ventilators for an Influenza Pandemic

2.1 Introduction

When a novel influenza virus emerges, it can spread quickly among people across a wide area due to lack of immunity and airborne transmission, causing great damage. The 1918 influenza pandemic caused an estimated 30-50 million deaths globally, of which 675 thousand were Americans, and the 1958 pandemic caused about 1-2 million deaths worldwide, of which about 70 thousand were in the U.S. [66]. The U.S. Department of Health and Human Services [65] estimates that 865 thousand people in the U.S. will be hospitalized in a moderate pandemic scenario (like 1958/68) and 9.9 million people in a severe pandemic scenario (like 1918). Several countermeasures, including vaccines, antivirals, and school closures, have been considered and studied to control and mitigate a pandemic.

When an influenza outbreak occurs, a surge of critical medical resources, e.g., mechanical ventilators, antivirals, personal protection equipment, etc., is required for treating sick people and containing the outbreak. In particular, mechanical ventilators have been a center of discussion since they are life-saving devices for people with severe acute respiratory failure. Decisions on how to stockpile such critical resources are challenging because of budget limits and the difficulty of predicting the timing and severity of the next pandemic. The mismatch between influenza-based demand for mechanical ventilators and existing capacity in intensive-care units (ICUs) has been

highlighted in the literature. Smetanin et al. [50] estimate the demand for ICU beds and ventilators in Manitoba, Canada. In addition, the study shows how a country, like Canada, may fall short of ventilators during pandemics more severe than the 2009 H1N1 pandemic. Stiff et al. [52] use mathematical modeling of disease spread to estimate the demand and investigate the possibility of demand-capacity mismatch for pediatric ICU beds in Canada during a pandemic. Ercole et al. [14] estimate the demand for critical medical care during an influenza pandemic, and they conclude that sentinel reporting and real-time modeling is critical for optimizing resource utilization in response to a pandemic.

Instead of stockpiling more mechanical ventilators, some researchers consider increasing current surge capacity by expanding existing ventilation capacity. Neyman and Irvin [38] show that a single ventilator with proper modification may offer sufficient ventilation for four 70-kg adults for 12 hours by testing it with lung simulators. Paladino et al. [45] also show the possibility of ventilating four 70-kg sheep for at least 12 hours, using a single ventilator with a four-limbed circuit.

In addition to existing hospital-owned ventilators, the Strategic National Stockpile (SNS) program of the U.S. Centers of Disease Control and Prevention (CDC) holds a large quantity of mechanical ventilators stored in strategically located warehouses for public health emergencies, such as an influenza outbreak, terrorist attack, etc. Once an emergency is announced and the SNS is deployed, the SNS ventilators will be delivered to states in need within days. However, the ventilators in the SNS may not be enough to meet the surge in demand during a severe public health emergency. As of May 25, 2006, the American Association for Respiratory Care (AARC) suggested that the CDC increase the SNS inventory by at least 5,000 to

10,000 ventilators, in addition to the existing 6,000 ventilators, for a severe influenza pandemic [1]. Currently, AARC partners with the CDC/SNS to provide training on the three kinds of ventilators in the SNS: LP-10, LTV-1200, and Uni-vent [2], so that these SNS ventilators will be well-used when delivered to local hospitals. Furthermore, states may receive federal funding to build and maintain their own ventilator supply for public health emergencies. Wilgis [69] quantitatively discusses the relative merits of stockpiling ventilators at a single location and then distributing them during an emergency versus distributing ventilators a prior to hospitals. The advantages of stockpiling at one site include maintaining accurate counts, ensuring timely repairs, optimizing allocation during a pandemic, etc., while the pros of distributing ventilators to hospitals ahead of time include improving hospital staff trouble-shooting skills, having ready-to-use ventilators, incurring no central warehouse cost, etc.

The notion of risk pooling is pervasive in insuring against financial losses and in supply chain management (see [51] for concrete examples). In the insurance industry, by pooling a large number of risk exposures, an insurer can take advantage of the fact that these potential liabilities are realized for a multitude of reasons. Hence, the insurer need not cover each risk separately, but rather reduce reserves required to cover the aggregate risk with high probability [5]. In the latter setting, and more specifically in inventory management, by aggregating stochastic demands we can reduce the total relative variation dramatically, especially when demands are negatively correlated, and, in turn, reduce necessary inventory for a given risk level [49]. Stockpiling ventilators centrally and then distributing them regionally as needed, e.g., as with SNS ventilators, is an example of risk pooling.

When the ventilator supply is limited, a natural question is how to allocate

available ventilators. The Ethics Subcommittee of the Advisory Committee to the Director of the CDC [15] suggests using a multi-principle allocation system to account for diverse moral considerations while allocating ventilators during a pandemic influenza. Powell et al. [46] develop an ethical framework and then use it to derive a set of ethical and clinical guidelines for allocating ventilators in the state of New York. In particular, the sequential organ failure assessment (SOFA) score is used in the guidelines to evaluate a patient's need for ventilation. The Minnesota Department of Health and the Texas Department of State Health Services also recommend that healthcare facilities assign limited ventilators to patients most likely to benefit by using the SOFA scoring table [36, 58]. Although the SOFA score is widely recommended for ventilator triage, it requires four laboratory measurements which may be impractical to obtain during an influenza pandemic. Grissom et al. [18] propose using a modified SOFA (MSOFA) score, which only requires one laboratory measurement by showing that the MSOFA score can predict mortality as well as the SOFA score.

According to a detailed literature review of strategies for managing and allocating scarce resources during mass casualty events [60], there are few studies conducting quantitative research on stockpiling ventilators. And, none of the studies reviewed in [60] consider managing limited resources across multiple communities to optimally match supply and demand under various pandemic attack-rate scenarios.

In this chapter we present a method that a state can use to stockpile a critical medical resource, such as mechanical ventilators, both centrally and regionally, for an influenza pandemic. In the state of Texas, the Department of State Health Services (DSHS) can stockpile ventilators centrally, and hospitals in the state's eight health service regions (HSRs; see Figure 2.1) also stockpile ventilators. We use the state of

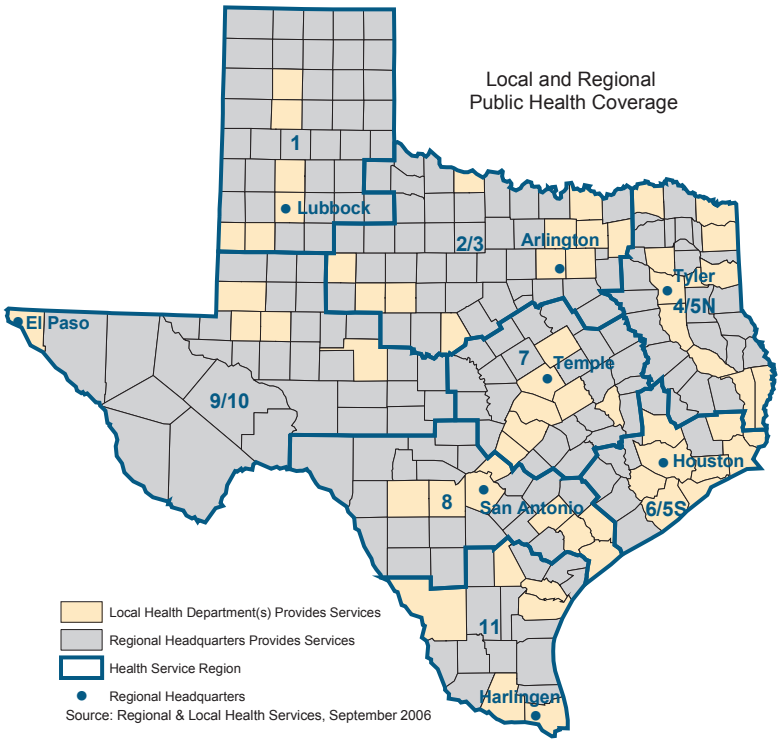


Figure 2.1: The eight health service regions in the state of Texas. Source: DSHS [56].

Texas as a case study to illustrate our approach. Our analysis is based on estimates of the peak number of patients that will require staffed ventilator beds during a pandemic for each of the eight HSRs in Texas [23]. By mapping hospitalizations of influenza-like illness (ILI) to ventilator demands, we obtain these estimates in the form of a multivariate normal distribution. The optimization model we develop for stockpiling also takes as input an upper bound on the expected shortfall of ventilators over the eight regions. We develop model variants that allow optimizing over both the central and regional stockpiles or allow taking either the central or regional stockpiles as input.

The organization of the remainder of this chapter is as follows. In Section 2.2, we describe our modeling framework, including estimating peak demands for ventilators, developing a stockpile model for quantitatively assessing the tradeoff between number of ventilators stockpiled and the associated risk, and presenting various uses of the model. We present algorithms to solve the stockpile model and detail how we overcame certain challenges in Section 2.3. In Section 2.4, we analyze stockpiling strategies under mild, moderate, and severe pandemic scenarios. We discuss the results and perform sensitivity analysis on the input parameters in Section 2.5.

2.2 Modeling Framework

We describe the methodology we use in this chapter as follows. First, in separate work [23], a Bayesian dynamic linear model (DLM) forecasts ILI hospitalizations and, in turn, peak demand for ventilators using multiple predictors. These forecasts are generated for each of the eight HSRs in Texas in the form of a multivariate normal distribution. Transforming forecasts for ILI hospitalizations to forecasts for peak ventilator demands requires three parameters, which we describe below. Then, a Monte Carlo sampling algorithm is employed to generate independent and identically distributed (i.i.d.) samples of peak demand for ventilators. Here, we need to specify the correlations between demands for ventilators in the eight HSRs. Third, a stochastic stockpile model takes as input the i.i.d. samples, existing central or regional stockpiles, and the proportion of ventilators sent from DSHS to regions that can be used. Then, the stockpile model computes the central and regional stockpiles needed to limit expected unmet demand (EUD) over the eight HSRs to a pre-specified threshold. Based on these central and regional stockpiles, we calculate the corresponding probability of shortfall of ventilators. Figure 2.2 shows a flowchart of our approach,

and we describe the details for each of these steps in the remainder of this section.

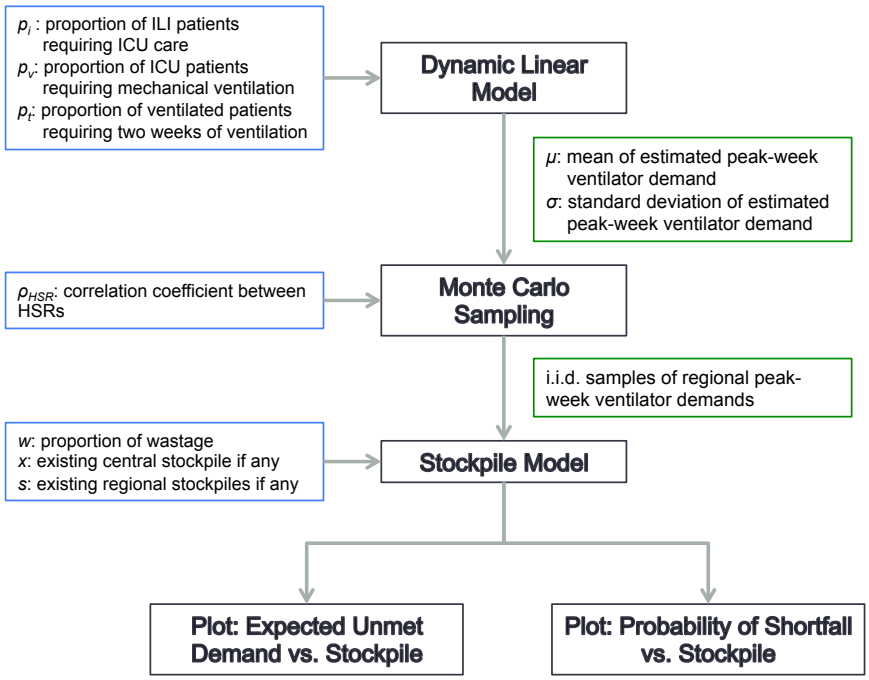


Figure 2.2: The steps and the associated inputs and outputs of the methodology. First, we use a DLM with three parameters to estimate peak-week ventilator demand for each region. Second, we employ a Monte Carlo sampling algorithm to generate a set of i.i.d. samples of regional ventilator demands. Third, we solve the stockpile model with proper inputs to generate a tradeoff curve between expected unmet demand and total stockpile as well as a tradeoff curve between probability of shortfall and total stockpile.

2.2.1 Demand Forecasting

We obtain estimates of peak-week ventilator demands based on forecasts for ILI hospitalizations. The hospitalization forecasts for each of the eight HSRs are represented as a multivariate normal distribution. The DLM [23] provides the means and variances of the marginal hospitalization distributions for each HSR; however,

it dose not capture the correlations between the regions. We estimate the region-to-region correlations by investigating historical data, which we discuss in further detail in Section 2.4. In order to estimate the peak demands for ventilators from the forecasted hospitalizations, we need three additional parameters: (i) p_i, the proportion of hospitalized patients who require ICU beds, (ii) p_v, the proportion of ICU patients who need ventilation treatment, and (iii) p_t, the proportion of ventilated patients who require two weeks of ventilation (under the assumption that a patient who requires a ventilator will need either one or two weeks of ventilation).

2.2.2 Model Assumptions

The models we formulate, and associated analysis we carry out, in this chapter are based on the following assumptions:

1. The model's geographic resolution is at the level of the eight health service regions as officially managed by DSHS. These regions represent aggregates of the state's 22 trauma service areas [59].

2. Ventilators are held in a central stockpile managed by DSHS and in regional stockpiles. Finer resolution, e.g., at the level of a county or individual hospital, is not modeled.

3. Within a region, a patient needing a ventilator is matched with a ventilator if one is available. Ventilators dedicated to a region cannot be shared outside that region. Patients needing a ventilator will not travel to another region to find one.

4. The modeling framework focuses on non-depletable ventilators and does not

consider consumable ventilator supplies or requisite staffing.

5. All ventilators are assumed to be both pediatric and adult capable.

6. Baseline demand, i.e., non-ILI patients who require ventilation, is neglected. The supply of ventilators in each region is best viewed as a supply that already has baseline demand from non-ILI patients subtracted.

7. Peak demand is modeled and different timing of peaks in different regions is neglected. Hence, we assume that DSHS cannot distribute a ventilator to one region and then subsequently retrieve it and redistribute the same ventilator to another region.

8. The correlation coefficient is assumed to be identical for each pair of regions.

2.2.3 Model Notation

We use the following notation in this chapter.

Indices and Sets

$r \in R$	:	health service regions
$\omega \in \Omega$	:	scenarios

Data and Parameters

μ_r	:	mean of estimated peak demand for ventilators in region r
σ_r	:	standard deviation of estimated peak demand for ventilators in region r
$d_r(\omega)$	:	peak demand for ventilators in region r under scenario ω
w_r	:	proportion of ventilators sent from DSHS to region r that cannot be used
L	:	upper limit on expected shortfall of ventilators summed over the regions

Decision Variables

x : number of ventilators stockpiled by DSHS

s_r : number of ventilators at hospitals in region r

$y_r(\omega)$: number of ventilators sent from DSHS to region r under scenario ω

The notation μ, σ, $d(\omega)$, w, s, and $y(\omega)$ represent the vector forms of μ_r, σ_r, $d_r(\omega)$, w_r, s_r, and $y_r(\omega)$; i.e., $\mu = (\mu_r)_{r\in R}$, $\sigma = (\sigma_r)_{r\in R}$, $d(\omega) = (d_r(\omega))_{r\in R}$, $w = (w_r)_{r\in R}$, $s = (s_r)_{r\in R}$, and $y(\omega) = (y_r(\omega))_{r\in R}$.

2.2.4 Model

We assume DSHS and hospitals in the eight HSRs need to decide the number of ventilators to stockpile before an influenza pandemic occurs. After an influenza pandemic begins, DSHS ships its stockpiled ventilators to the eight HSRs according to the realized peak demand for ventilators in each region. We also include a notion of wastage to capture the fact that centrally stockpiled ventilators may be less effective once they are distributed to a region than ventilators that have been stored, maintained, and operated locally in that region. Wastage may arise for multiple reasons: it takes time to ship ventilators to the region and more time to put the ventilators in the hands of those who need them; staff in the region may not be fully trained on the type of ventilators that were shipped; and, there can be mismatches in specific required types of supplies (e.g., filters).

In order to represent the decision-making process, we construct a two-stage stochastic program. Decisions x and/or s must be selected prior to observing the demand for ventilators. The decision to ship ventilators to region r is captured by decision variable $y_r(\omega)$. This decision is made after observing the demand realization under scenario ω. In addition, if $y_r(\omega)$ ventilators are shipped, then $w_r y_r(\omega)$ represents

the number of ventilators wasted so that only $(1 - w_r)y_r(\omega)$ ventilators can be used in region r. Hence, the model seeks a balance between (i) the flexibility permitted by holding ventilators centrally so that they can be distributed to where they are needed most and (ii) the fact that locally held ventilators are more effective than those shipped from DSHS after a pandemic begins.

The stockpile model is as follows:

$$z^*(L) = \min_{x,s,y} \quad x + \sum_{r \in R} s_r \tag{2.1a}$$

$$\text{s.t.} \quad \sum_{r \in R} y_r(\omega) \leq x, \forall \omega \in \Omega \tag{2.1b}$$

$$\mathbb{E}_\omega \left[\sum_{r \in R} \left[[d_r(\omega) - s_r]^+ - (1 - w_r)y_r(\omega) \right]^+ \right] \leq L \tag{2.1c}$$

$$x \geq 0, s_r \geq 0, y_r(\omega) \geq 0, \forall r \in R, \omega \in \Omega. \tag{2.1d}$$

The objective function in (2.1a) is the total stockpile of central and local ventilators, which we want to minimize. Constraint (2.1b) says that the total number of ventilators distributed from DSHS to the regions cannot exceed the number of ventilators stockpiled by DSHS. Moreover, $[d_r(\omega) - s_r]^+ = \max\{d_r(\omega) - s_r, 0\}$ represents the amount by which peak demand for ventilators exceeds existing supply in region r under scenario ω, and $\sum_{r \in R} \left[[d_r(\omega) - s_r]^+ - (1 - w_r)y_r(\omega) \right]^+$ represents the total shortfall of ventilators statewide after distributing the central stockpile under scenario ω. Then constraint (2.1c) ensures that the expected shortfall of ventilators over the eight HSRs does not exceed the limit, L. Constraint (2.1d) enforces non-negativity for each decision variable. Note that $d(\omega)$ and w are input data and $y(\omega)$ is a decision

variable. By specifying the values of x, s, or neither, three variations of the model can address the stockpile problem from different perspectives: (i) given existing regional stockpiles, optimize the number of centrally held ventilators; (ii) given an existing central stockpile, optimize the number of regionally held ventilators; and, (iii) jointly optimize the central and regional stockpiles, assessing the advantages of stockpiling ventilators centrally versus regionally. Model (2.1) is stated in the form of variant (iii), but we can formulate variant (i) or (ii) by fixing decision variables s or x, respectively, to pre-specified values. We now explain the variants of the model for each of these three perspectives.

First version of the model: Input: s. Output: x.

Our first variant of the model takes the regional demands ($d(\omega)$), existing regional stockpiles (s), and the limit on EUD (L) as input and yields as output the smallest number of ventilators that DSHS should stockpile (x) to ensure the limit on EUD is satisfied. This model can help DSHS decide whether an existing, or planned, stockpile of ventilators is adequate under various demand scenarios, e.g., for a severe pandemic like that in 1918 or a mild pandemic like that in 2009.

Second version of the model. Input: x. Output: s.

The second variant of the model takes regional demands ($d(\omega)$), the DSHS stockpile (x), and the limit on EUD (L) as input and yields as output the stockpiles for the regions (s) to hold to ensure the limit on EUD is satisfied. This model minimizes the total stockpile across all regions, and can yield stockpiling recommendations for the regions given the current, or planned, DSHS stockpile.

Third version of the model. Output: x and s.

The third variant of the model seeks the optimal number of ventilators to

hold centrally (x) and in the eight HSRs (s), given regional demands ($d(\omega)$) and the limit on EUD (L). In the model we minimize the sum of centrally stockpiled and regionally stockpiled ventilators. The purpose of the model is to help DSHS assess the advantages of stockpiling ventilators centrally versus regionally.

2.3 Solution Method

This section presents the solution method we use to solve the optimization models described in Section 2.2. We develop a Monte Carlo sampling-based approximation to solve model (2.1) for the following reasons. The summed shortfall of ventilators, i.e., $\sum_{r\in R}\left[[d_r(\omega)-s_r]^+-(1-w_r)y_r\right]^+$, is a non-standard random variable due to the two positive-part operators within the summation, although $d(\omega)$ has the form of a multivariate normal distribution. More importantly, the decision variables, $y(\omega)$, representing shipments to the eight regions adapt to the demand realization under scenario ω, increase the model's complexity. In what follows, we first describe the scheme for generating i.i.d. samples from a multivariate normal distribution. Then, we explain the method for solving model (2.1) approximately with the samples, the associated technical hurdles we encountered, and how we overcame those hurdles.

2.3.1 Sampling from a Multivariate Normal Distribution

We use Algorithm 1 to generate n i.i.d. samples from a multivariate normal distribution with mean vector μ and covariance matrix Σ. The first step of the algorithm is to compute the Cholesky factorization C of Σ, i.e., C is a lower triangular matrix with $\Sigma = CC^\top$. The algorithm then repeats the following process n times: produce $|R|$ independent standard normal variates and then scale and shift those to produce an $|R|$-dimensional random vector with mean μ and covariance matrix Σ.

This is a standard algorithm; see, for example, Devroye's text [13].

Algorithm 1 Monte Carlo Method for Generating Observations from a Multivariate Normal

Input: $|R|$-dimensional mean vector μ, $|R| \times |R|$ dimensional covariance matrix Σ, sample size n

Output: n i.i.d. observations, $d^1, d^2, \ldots, d^n$, of multivariate normal with mean μ and covariance Σ

Let $K = |R|$ (the number of health service regions)
Compute Choleksy factorization of Σ to obtain lower triangular matrix

$$C = \begin{pmatrix} C_{11} & 0 & \cdots & 0 \\ C_{21} & C_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ C_{K1} & C_{K2} & \cdots & C_{KK} \end{pmatrix}$$

for $i = 1$ to n **do**
  Generate $Z_k \sim N(0,1)$, $k = 1, 2, \ldots, K$
  for $k = 1$ to K **do**
    $d_k^i = \mu_k + \sum_{k'=1}^{k} C_{kk'} Z_{k'}$
  end for
end for

return $d^i = (d_1^i, d_2^i, \ldots, d_K^i)$, $i = 1, 2, \ldots, n$

2.3.2 Approximate Stochastic Model

Let $i = 1, 2, \ldots, n$ index the sample scenarios. Our sampling-based variant of model (2.1) is as follows:

$$z_n^*(L) = \min_{x,s,y,u,v} \quad x + \sum_{r \in R} s_r \tag{2.2a}$$

$$\text{s.t.} \quad \sum_{r \in R} y_r^i \leq x, \forall i = 1, 2, \ldots, n \tag{2.2b}$$

$$u_r^i \geq d_r^i - s_r, \forall r \in R, i = 1, 2, \ldots, n \tag{2.2c}$$

$$v_r^i \geq u_r^i - (1 - w_r) y_r^i, \forall r \in R, i = 1, 2, \ldots, n \tag{2.2d}$$

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{r \in R} v_r^i \leq L \tag{2.2e}$$

$$x \geq 0, s_r \geq 0, y_r^i \geq 0, u_r^i \geq 0, v_r^i \geq 0,$$
$$\forall r \in R, i = 1, 2, \ldots, n. \tag{2.2f}$$

Again, the objective function in (2.2a) is the total number of central and regional ventilators, which we want to minimize. Constraint (2.2b) is analogous to constraint (2.1b), where we add index i to variable y_r because shipments from the central stockpile to the regions occur after observing the demand realization. In constraints (2.2c) and (2.2d), $d^i = (d_r^i)_{r \in R}$, $i = 1, 2, \ldots, n$, are the samples of regional ventilator demands obtained from Algorithm 1, and $(1 - w_r)$ is the proportion of centrally held ventilators sent to region r that can be used. These two constraints

take care of the two positive-part operators by using two new decision variables: u_r^i and v_r^i. Given that these variables capture these positive parts, constraint (2.2e) is analogous to constraint (2.1c), and constraint (2.2f) again captures non-negativity of all decision variables. While we state models (2.1) and (2.2) for a fixed value of L, we can view this as a bi-criteria model in which we can explore the tradeoff between the cost of the total stockpile (which we assume to be proportional to the number of ventilators) and the limit on expected shortfall (L).

2.3.3 Technical Hurdles and Solutions

When using model (2.2) to solve to model (2.1) approximately, there are still several technical hurdles we face, including excessive solution time, ragged stockpile solutions as we range the limit on EUD (L), and a concern about over-optimizing with respect to the n i.i.d. demand scenarios. We describe each issue in detail and how we overcame these hurdles in the rest of this section. Figure 2.3 shows the three steps and the associated solution methods.

2.3.3.1 Issue 1: Excessive Solution Time

In model (2.2), we seek to solve a bi-criteria model, capturing the tradeoff between the number of ventilators required and risk level L. However, solving each instance of model (2.2) for a specific value of L can take significant time. We use the fact that the optimal value, i.e., the total number of ventilators, is a convex function of L to reduce the number of instances we must solve of this parametric model in Step 1 of Figure 2.3. We use Algorithm 2 to approximate the tradeoff curve between the total stockpile and L.

Assume that we have m instances of model (2.2) to solve, which we sort in

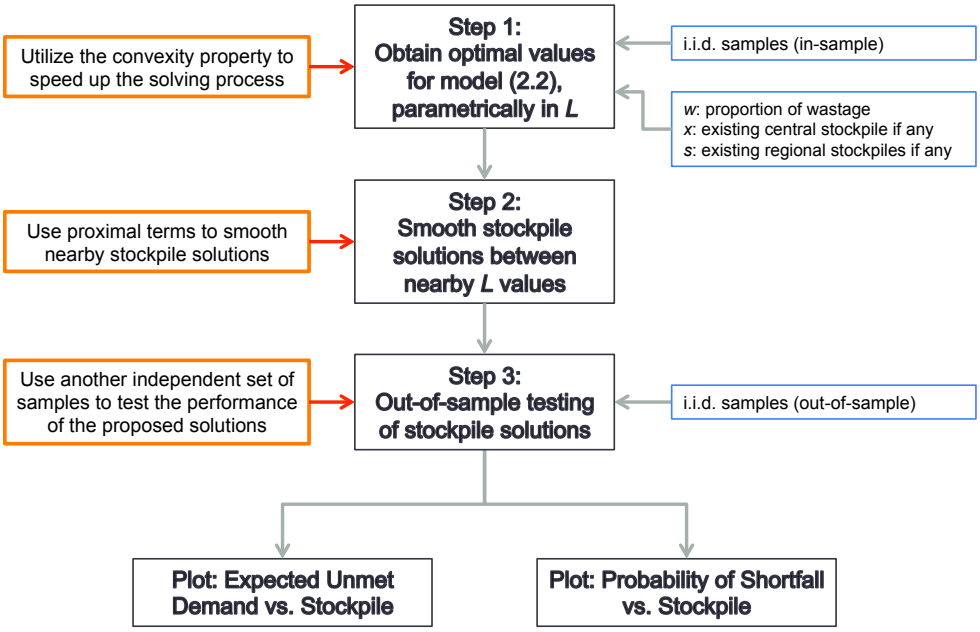


Figure 2.3: Solution procedure associated with "Stockpile Model" step in Figure 2.2. First, we use convexity of the optimal value of model (2.2) in L to reduce the required computation effort. Second, we employ a variant of model (2.2) with proximal terms to smooth stockpile solutions between nearby L values. Finally, we use another independent set of i.i.d. samples of regional peak demands for ventilators to test the performance of the proposed stockpile solutions and provide as output the two tradeoff curves.

ascending order of their L values so that instance 1 has the smallest value of L (L^1), and instance m has the largest value of L (L^m). For instance j with L^j, we let z_n^{j*}, $\underline{z}_n^{j*}$, and $\overline{z}_n^{j*}$ denote its optimal value, lower bound, and upper bound, respectively. We use δ^j to denote the gap between $\underline{z}_n^{j*}$ and $\overline{z}_n^{j*}$, i.e., $\delta^j = \overline{z}_n^{j*} - \underline{z}_n^{j*}$, and ϵ_{z_n} to denote a tolerance on this gap.

The first step of Algorithm 2 is to solve the two boundary instances, instances 1 and m, to optimality, obtain z_n^{1*} and z_n^{m*}, and set both their lower bounds and upper bounds to be their optimal values, i.e., $\underline{z}_n^1 = \overline{z}_n^1 = z_n^{1*}$ and $\underline{z}_n^m = \overline{z}_n^m = z_n^{m*}$. We add these two optimal values with their L values to a set of frontier points, denoted F, in Step 2.

In Step 3, for each unsolved instance j the algorithm calculates its lower bound ($\underline{z}_n^j$), which is the maximum of two first-order Taylor approximations: one from the nearest solved instance with smaller value of L, with its shadow price of constraint (2.2e) as the slope, and the other from the nearest solved instance with larger value of L. The shadow price of constraint (2.2e) of each solved instance serves as a subgradient for the convex tradeoff curve.

In Step 4, for each unsolved instance j the algorithm calculates its upper bound ($\overline{z}_n^j$), which is the interpolation of the nearest adjacent optimal values with smaller and larger values of L. Due to convexity of the optimal number of ventilators as a function of L, the corresponding chord lies above the efficient frontier, providing an upper bound.

In Steps 5 and 6, the algorithm computes δ^j for each instance and obtains the maximum gap, denoted δ_{max}. If δ_{max} is less than the pre-specified ϵ_{z_n}, the algorithm outputs the frontier points in F for approximating the tradeoff curve. If not, the

algorithm finds an instance, denoted k, with the maximum gap, i.e., $\delta^k = \delta_{max}$, solves instance k, adds the optimal value and the associated value of L to F, and repeats Steps 3 to 6.

Algorithm 2 Use the convexity of $z_n^*(L)$ to solve parametric model (2.2)

Input: (i) a set of instances of model (2.2) with different values of L in ascending order, i.e., $L^1 < L^2 < \cdots < L^m$, and (ii) tolerance in optimality, ϵ_{z_n}
Output: a set of frontier points of the tradeoff curve, F

Let $F = \emptyset$ (used for recording frontier points of the tradeoff curve)

Step 1: solve instances of model (2.2) with L^1 and L^m, obtain z_n^{1*} and z_n^{m*}, and set $\underline{z}_n^1 = \overline{z}_n^1 = z_n^{1*}$ and $\underline{z}_n^m = \overline{z}_n^m = z_n^{m*}$
Step 2: add (L^1, z_n^{1*}) and (L^m, z_n^{m*}) to F
Step 3: update $\underline{z}_n^j$, $j = 1, 2, \ldots, m$ via Taylor series approximation using shadow prices of constraint (2.2e)
Step 4: update $\overline{z}_n^j$, $j = 1, 2, \ldots, m$ by interpolation
Step 5: $\delta^j = \overline{z}_n^j - \underline{z}_n^j$, $j = 1, 2, \ldots, m$
Step 6: $\delta_{max} = \max_j\{\delta^j\}$

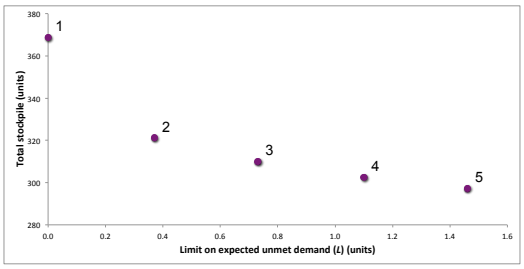
while $\delta_{max} > \epsilon_{z_n}$ **do**
(i) let $k \in \arg\max_j\{\delta^j\}$
(ii) solve instance of model (2.2) with L^k and obtain z_n^{k*} and shadow price of constraint (2.2e)
(iii) add (L^k, z_n^{k*}) to F
(iv) repeat Steps 3 to 6
end while

return F

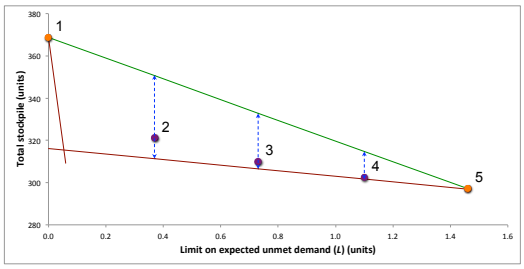
Figure 2.4 illustrates how Algorithm 2 works with a small example in which we have five instances of model (2.2) with different values of L, as nodes 1-5 in Figure 2.4(a). The values of L are 0, 0.37, 0.73, 1.10, and 1.46, with units of patients with unsatisfied demand for a ventilator. First, the algorithm solves instances associated with nodes 1 and 5 and adds them to the set of frontier points, i.e., $(0, 368.76)$

and $(1.46, 297.08)$, where the latter number in the ordered pair is the total number of stockpiled ventilators. Next, the algorithm calculates a lower bound and upper bound for each of the instances associated with nodes 2-4, as shown in Figure 2.4(b). The shadow price of constraint (2.2e) is -986.49 for the instance associated with node 1 and -13.53 for the instance associated with node 5, and we use these to form first-order Taylor series approximations at nodes 1 and 5 as the figure depicts. We obtain the lower bounds for the instances associated with nodes 2-4 as 311.41, 306.54, and 301.68. We also have upper bounds as 350.59, 332.92, and 314.75 by interpolating the solved instances associated with nodes 1 and 5. The instance associated with node 2 has the largest optimality gap 39.18 (= 350.59 - 311.41). Assuming this gap is larger than the pre-specified threshold, the algorithm then solves the instance associated with node 2 and adds it to the set of frontier points, i.e., (0.37, 321.32). The shadow price of constraint (2.2e) is -41.11 for the instance associated with node 2. We then update the lower and upper bounds for the instances associated with nodes 3-4, as shown in Figure 2.4(c). The optimality gaps of instances associated with nodes 3-4 are now 6.77 and 3.41. If the maximum gap, i.e., 6.77 for the instance associated with node 3, is smaller than the threshold, the algorithm outputs the current frontier points, i.e., nodes 1, 2, and 5, for approximating the tradeoff curve, as shown in Figure 2.4(d). If not, the algorithm continues to solve next instance with the maximum gap and updates the lower and upper bounds.

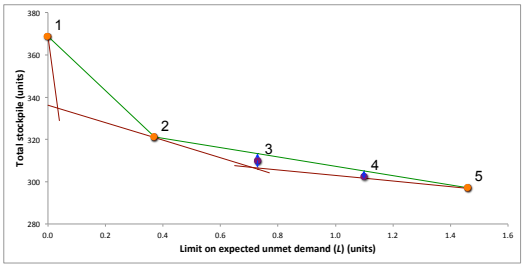
When simulating ventilator stockpiling under mild, moderate, and severe pandemic scenarios in Section 2.5, we set the number of discretized L values to be 1,000 with a minimum of 0 and a maximum of the largest total ventilator demand sampled, denoted d_{max}, i.e., $L^0 = 0$ and $L^{1000} = d_{max}$. We also set $\epsilon_{z_n} = 1$ as default.



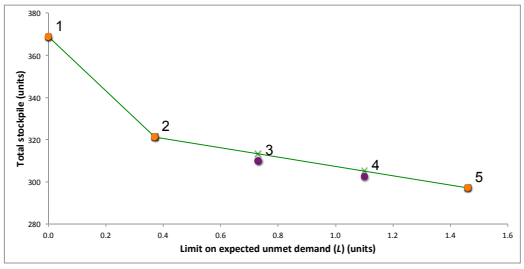
(a) Optimal values for nodes 1 to 5



(b) Solved nodes: 1 and 5



(c) Solved nodes: 1, 2, and 5



(d) Approximation output

Figure 2.4: Illustration of using Algorithm 2 to accelerate solution of the parametric model (2.2). First, we solve two boundary instances, nodes 1 and 5. Next, we calculate the lower bound, upper bound, and optimality gap at each value of L, as shown in part (b) of the figure. Assuming that node 2 has the maximum gap and that gap is larger than the pre-specified threshold, we then solve the instance at node 2 and update the lower bound, upper bound, and optimality gap at each value of L, as shown in part (c) of the figure. If the optimality gaps of nodes 3 and 4 are both smaller than the threshold, we output the optimal values and L values of nodes 1, 2, and 5 for approximating the tradeoff curve between total stockpile and L, as shown in part (d) of the figure. If not, we continue to solve the instance at node 3 because it has the maximum gap among all unsolved instances.

Algorithm 2 can help us significantly accelerate the process of solving the parametric model (2.2) over a brute force approach of simply solving the stochastic linear program at each value of L. For example, when we simulate the mild pandemic scenario with the default setting, we obtain 21 frontier points for approximating the tradeoff curve, which means we save about 98% (= 979/1000) of the computational effort relative to the brute force approach of solving 1,000 model instances.

2.3.3.2 Issue 2: Optimal Solutions between nearby L values are Ragged

The objective function of model (2.2) can be relatively flat in the neighborhood of optimal solutions, and so there may be multiple optimal solutions for a specified L value. As a result, when numerically solving the parametric model (2.2), we may obtain optimal solutions at nearby values of L that are unnecessarily ragged. This variable nature of the optimal central and regional stockpiles, as a function of L, is undesirable. So, we use proximal (i.e., target) terms in an auxiliary linear program in Step 2 of Figure 2.3 to smooth the stockpile allocation as we parametrically change

L. The auxiliary model is as follows:

$$\min_{x,s,\overline{s},y,u,v} \quad \sum_{r\in R} |s_r - s_{r,target}| + |s_r - \overline{s}| \tag{2.3a}$$

$$\text{s.t.} \quad \text{(2.2b), (2.2c), (2.2d), (2.2e), (2.2f)}$$

$$\overline{s} = \frac{\sum_{r\in R} s_r}{|R|} \tag{2.3b}$$

$$x + \sum_{r\in R} s_r \leq (1 + \epsilon_{target}) \cdot z_n^*(L) \tag{2.3c}$$

$$x \leq x_{target}. \tag{2.3d}$$

We use x_{target} and $s_{r,target}$, obtained by solving model (2.3) at a nearby value of L, to denote the targets for central and regional stockpiles. The objective function in (2.3a) includes two terms. The former term computes the one norm of the difference between the regional stockpiles and their targets, and the latter term computes the absolute deviation of the regional stockpiles from their average. Minimizing this objective function helps us obtain smooth regional stockpiles between nearby L values (by the former term) and balance regional stockpiles for each individual L value (by the latter term). In addition to constraints (2.2b)-(2.2f), we have three other constraints to satisfy. Constraint (2.3b) calculates the average regional stockpile, denoted $\overline{s}$. Constraint (2.3c) ensures that the total stockpile is at most ϵ_{target} more than the optimal total stockpile we obtain from Algorithm 2. We set $\epsilon_{target} = 0.001$ as default. Constraint (2.3d) requires the central stockpile to be no more than its target.

We use Algorithm 3 to smooth the stockpile allocations of nearby frontier points obtained from Algorithm 2. For an instance of model (2.3) with L^l, we use $(\hat{x}^l, \hat{s}^l)$ to denote the near-optimal stockpile allocation obtained from Algorithm 3. Algorithm 3 requires that we specify initial values of x_{target} and $s_{target}(= (s_{r,target})_{r \in R})$. We prefer to stockpile ventilators centrally rather than regionally, so that we set initial x_{target} to be d_{max}, and s_{target} to be $\mathbf{0}$.

Assume we have a total of q instances of model (2.3) to solve, which we sort in ascending order of their L values so that instance 1 has the smallest value of L (L^1), and instance q has the largest value of L (L^q). First, we set $x_{target} = d_{max}$ and $s_{target} = \mathbf{0}$. Then, Algorithm 3 solves the instance of model (2.3) with L^1, obtains $(\hat{x}^1, \hat{s}^1)$, and updates x_{target} to be $\hat{x}^1$ and s_{target} to be $\hat{s}^1$. Next, the algorithm solves the problem with L^2 and updates x_{target} and s_{target}. This process repeats until all instances are solved.

Algorithm 3 Smooth optimal solutions between nearby L values

Input: (i) a set of instances of model (2.3) with L values in ascending order, i.e., $L^1 < L^2 < \cdots < L^q$ and their optimal total stockpile, $z_n^{1*}, z_n^{2*}, \ldots, z_n^{q*}$, (ii) tolerance in optimal total stockpile, ϵ_{target}, and (iii) maximum total demand for ventilators sampled, d_{max}

Output: smooth near-optimal solutions between nearby L values for the instances of model (2.2), $(\hat{x}^1, \hat{s}^1), (\hat{x}^2, \hat{s}^2), \ldots, (\hat{x}^q, \hat{s}^q)$

Set $x_{target} = d_{max}$ and $s_{target} = \mathbf{0}$

for $l = 1$ to q **do**
(i) solve instance of model (2.3) with L^l and obtain a near-optimal stockpile allocation $(\hat{x}^l, \hat{s}^l)$ for the instance of model (2.2) with L^l.
(ii) set $x_{target} = \hat{x}^l$ and $s_{target} = \hat{s}^l$
end for

return $(\hat{x}^1, \hat{s}^1), (\hat{x}^2, \hat{s}^2), \ldots, (\hat{x}^q, \hat{s}^q)$

2.3.3.3 Issue 3: Assessing Solution Quality

Suppose we have used Algorithms 2 and 3 to obtain a near-optimal total stockpile, $(\hat{x}^l, \hat{s}^l)$, for $L = L^l$. For this near-optimal solution, constraint (2.2e) may not be tight, i.e., the left-hand side of constraint (2.2e) evaluated at $(\hat{x}^l, \hat{s}^l)$ may be less than the limit, L^l. We use the following model to obtain the EUD for each smoothed frontier point.

$$L_n^*(\hat{x}, \hat{s}) = \min_{y,u,v,L} \quad L \tag{2.4a}$$

$$\text{s.t.} \quad \text{(2.2b), (2.2c), (2.2d), (2.2e)}$$

$$y_r^i \geq 0, u_r^i \geq 0, v_r^i \geq 0, \forall r \in R, i = 1, 2, \ldots, n. \tag{2.4b}$$

Note that x and s now are inputs, which are set to $\hat{x}$ and $\hat{s}$ obtained from Algorithm 3, and the decision variables are y, u, v, and L. We keep constraints (2.2b)-(2.2e), except that we replace the right-hand side of constraint (2.2b) with $\hat{x}$ and the right-hand side of constraint (2.2c) with $d_r^i - \hat{s}_r$, and we modify constraint (2.2f) to constraint (2.4b) since x and s are fixed as input. By solving model (2.4), we obtain as its optimal value the EUD for stockpile allocation $(\hat{x}^l, \hat{s}^l)$, $l = 1, 2, \ldots, q$.

Furthermore, a key driver in this process is the corresponding expected unmet demand (EUD), i.e., the left-hand side of constraint (2.1c). Because we solve a sample average approximation in model (2.2) for a specific set of n sampled scenarios, called in-sample scenarios, the actual EUD of $(\hat{x}^l, \hat{s}^l)$ will differ from the EUD we have obtained. In particular, the concern is that because the stockpiling decisions, x and s, are "tuned" to the in-sample scenarios used in model (2.2), the true EUD may

be larger than the EUD we obtained. We can quantify this by fixing the x and s decisions, generating another independent set of n scenarios, called out-of-sample scenarios, and re-estimating EUD, using model (2.4); i.e., re-estimating the left-hand side of constraint (2.1c). We emphasize that the n out-of-sample scenarios we use now are independent of those used in Algorithms 2 and 3 with models (2.2) and (2.3) to find x and s.

We use the mild pandemic scenario, which we describe in detail in Section 2.4, to illustrate the quality of solutions obtained by our sampling-based solution method. Figure 2.5 shows two tradeoff curves that correspond to in-sample and out-of-sample scenarios. Based on the in-sample set of scenarios, Algorithms 2 and 3 generate 21 frontier points and their stockpile allocations, depicted by the blue tradeoff curve in the figure. We then use model (2.4) and out-of-sample scenarios to estimate EUD for these 21 stockpile solutions, depicted by the red line in the figure. For a fixed stockpile, the red curve tends to shift slightly to the right due to the over-optimization of the stockpiling decisions to the in-sample set of scenarios. The difference between these EUDs has an average of 0.7, a standard deviation of 0.75, and a maximum of 1.75 (all in units of ventilators), which suggests that the recommended stockpile solutions perform well. Moreover, if we allow zero tolerance in optimal total stockpile in Algorithm 2, i.e., $\epsilon_{z_n} = 0$, we obtain 315 frontier points with similar solution quality but with much more computing time, which suggests $\epsilon_{z_n} = 1$ works well. We use the tradeoff curve based on the out-of-sample scenarios as our best estimate for the true tradeoff between total stockpile and EUD in Section 2.5.

The results in Figure 2.5, are based on a single set of $n = 1,000$ i.i.d. realizations of the demand vector. The second column in Table 2.1 shows the numerical

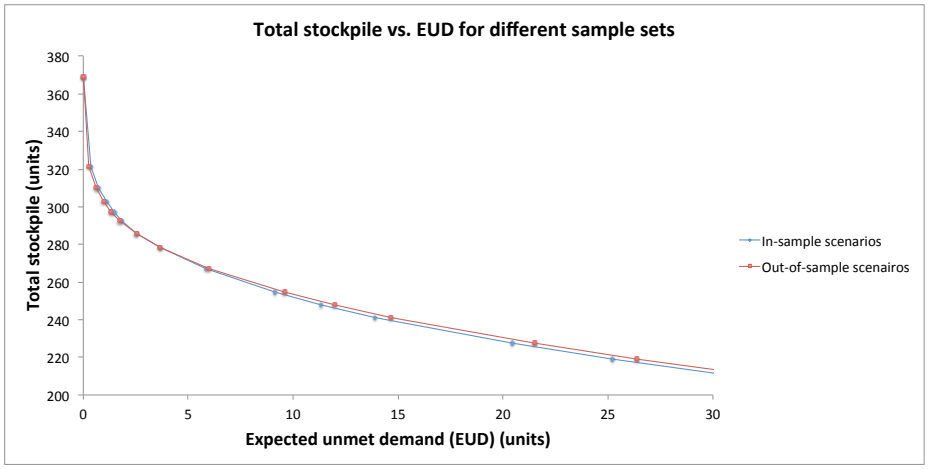


Figure 2.5: Assessing solution quality by estimating expected unmet demand (EUD) with another set of samples. After obtaining the frontier points and the associated stockpile allocations for the tradeoff curve based on $n = 1,000$ in-sample scenarios, we compute $L_n^*(\hat{x}^l, \hat{s}^l)$ from model (2.4) using $n = 1,000$ out-of-sample scenarios. The blue curve is the performance of these frontier points on the in-sample scenarios and the red curve is on the out-of-sample scenarios. We see that these frontier points have similar performance on both sample sets with 1.75 being the maximum absolute difference of EUD, which suggests a high quality of these sampling-based stockpile solutions.

EUD values for the 21 frontier points of the blue tradeoff curve in Figure 2.5, i.e., the solution to model (2.4) for these 21 frontier points on the in-sample set. To understand the variability associated with $n = 1,000$ out-of-sample scenarios (the red curve in Figure 2.5), we sample 30 independent out-of-samples sets of $n = 1,000$ scenarios. The third column shows a sample mean of thirty observations of $L_n^*(\hat{x}^l, \hat{s}^l)$ for each $l = 1, 2, \ldots, q$, where $q = 21$ and $n = 1,000$. The fourth column shows a corresponding 95% confidence interval halfwidth for $\mathbb{E}L_n^*(\hat{x}^l, \hat{s}^l)$. We can see that the difference between the EUD on the in-sample scenarios and the average EUD is less than 0.5 and the half-width is less than 1 (all in units of ventilators) for all 21 frontier points, which together show the high quality of the recommended stockpile solutions.

Furthermore, when presenting the results of the tradeoff curve based on a single out-of-sample set of scenarios in Section 2.5, we calculate a stockpile solution for each integral EUD value by interpolating the frontier points of the tradeoff curve. An integral EUD is a conservative performance measure of the associated interpolated stockpile solution on the out-of-sample scenarios due to the convexity of the optimal value on $(\hat{x}, \hat{s})$ in model (2.4). For example, under the mild scenario we obtain the recommended total stockpile of 271.71 for EUD of 5 from two frontier points (3.67, 278.22) and (5.99, 266.86), as shown in subsequent Figure 2.6(a).

2.4 Estimating Demand for Ventilators under Three Pandemic Scenarios

In this section, we describe how we estimate peak demands and other model parameters for three different pandemic influenza scenarios: mild, moderate, and severe.

Table 2.1: Assessing solution quality by estimating expected unmet demand (EUD) via 30 i.i.d. observations of $L_n^*(\hat{x}^l, \hat{s}^l)$ from model (2.4) with $n = 1,000$. The second column shows the EUD values for the 21 frontier points on the in-sample scenarios; i.e., the second column provides the numerical values of EUD for the blue curve in Figure 2.5. The third column shows a sample mean of thirty observations of $L_n^*(\hat{x}^l, \hat{s}^l)$ for each $l = 1, 2, \ldots, 21$; i.e., the third column provides a sample mean estimate that corresponds to the single replication (of 1,000 scenarios) reported in the red curve of Figure 2.5. The fourth column shows a corresponding 95% confidence interval halfwidth for $\mathbb{E}L_n^*(\hat{x}^l, \hat{s}^l)$. We can see that the difference between the EUD of in-sample scenarios and the average EUD is less than 0.5 and the half-width is less than 1 (all in units of ventilators) for all 21 frontier points, which together show the high quality of the recommended stockpile solutions.

Frontier point	EUD for the in-sample set (units)	Average EUD of 30 out-of-sample sets (units)	Half-width of 95% confidence interval (units)
1	0.00	0.02	0.01
2	0.37	0.36	0.03
3	0.73	0.72	0.04
4	1.10	1.08	0.06
5	1.46	1.44	0.06
6	1.83	1.80	0.07
7	2.56	2.51	0.08
8	3.65	3.56	0.09
9	5.85	5.68	0.13
10	9.14	9.04	0.17
11	11.33	11.31	0.20
12	13.89	13.92	0.23
13	20.46	20.59	0.30
14	25.21	25.41	0.33
15	31.43	31.65	0.38
16	39.83	40.05	0.42
17	51.89	52.09	0.48
18	70.16	70.29	0.54
19	115.47	115.43	0.60
20	229.49	229.32	0.62
21	230.04	229.88	0.62

2.4.1 Estimating Peak Ventilator Demand for the Mild Scenario

Our stockpiling analysis requires as input the forecasts for the distribution of peak-week ventilator demand across the eight HSRs. In separate work [23], we estimate distributions for weekly hospitalizations under a mild pandemic scenario via a DLM by using April-December 2009 hospital discharge data for the state of Texas. This forecasting model uses hospitalizations in one week to predict hospitalizations in the next week, and allows us to capture temporal correlations in hospitalizations for H1N1 influenza. We apply the DLM forecasting model separately to each of the eight HSRs to obtain forecasts for ILI hospitalizations for April-December 2009. Using historical data, we estimate the pairwise correlations for ILI hospitalizations between regions. The data suggest that these pairwise correlations do not differ significantly among the 28 ($=C_2^8$) pairs of HSRs, and so we assume this correlation coefficient to be the same for each pair with a value of $\rho_{HSR} = 0.70$ based on the data. To check that this value is reasonable, we also estimate pairwise correlations among HSRs using 2002-2008 hospital discharge data for seasonal influenza in Texas, using the same 2002-2008 data restricted to peak-weeks, and using the 2002-2008 data restricted to intensive-care units. Our analysis shows that using a single pairwise correlation of $\rho_{HSR} = 0.70$ also appears consistent with these three sets of data.

The output of the DLM model combined with the estimate of ρ_{HSR} provide a multivariate normal probability distribution for hospitalizations across the eight HSRs for each week of the planning horizon. We then map this distributional forecast for hospitalizations to a distributional forecast for peak-week ventilator demand. This requires four additional parameters: (i) p_i, the proportion of ILI patients requiring ICU care, (ii) p_v, the proportion of ICU patients requiring mechanical ventilation, (iii)

p_t, the proportion of ventilated patients requiring two-week ventilation, and (iv) the correlation between ILI hospital admissions in consecutive weeks. We now describe how we estimate, or approximate, these parameters.

(i) Proportion of ILI Patients Requiring ICU Care (p_i):

From the 2009 ILI hospitalization data, we estimate that the proportion of hospitalized patients requiring ICU care is about 18% for the peak week. DSHS H1N1 hospitalization reports for October-December 2009 [55] indicate that 23% of the 2,030 confirmed H1N1 hospital admissions in Texas required ICU care. For moderate and severe planning scenarios, the U.S. Homeland Security Council (HSC) [68] uses an ICU proportion of 15% for the overall pandemic and a proportion of 25.7% for the peak week. For seasonal flu, the CDC's FLUSURGE 2.0 [71, 72] uses a default proportion of 15% of admitted influenza patients requiring ICU care, and values near 15% are also used by the U.S. Department of Health and Human Services (HHS) [65]. (See Table 3.) The Texas hospital discharge data we analyze suggest that the proportion of patients admitted to the ICU increases at the peak week. This is consistent with the larger peak-week values for moderate and severe planning scenarios put forward by the HSC. We use a value of 20% for p_i for the mild scenario and increase this to 25% for the moderate and severe scenarios.

(ii) Proportion of ICU Patients Requiring Mechanical Ventilation (p_v):

The CDC's FLUSURGE 2.0 [71] uses a default value of 50% for the proportion of ICU patients requiring mechanical ventilation, and the HSC [68] uses this same value for both the overall pandemic and the peak week. This estimate is based on seasonal influenza. We also use a default value of 50% for p_v, and later run a sensitivity analysis ranging this value up to 67%.

(iii) Proportion of Ventilated Patients Requiring Two-Week Ventilation (p_t):

The CDC's FLUSURGE 2.0 [71] uses a default value of 10 days for mechanical ventilation of a patient with ILI. In our model, with weekly time resolution, this corresponds to 57% of ventilated patients requiring one week and 43% requiring a second week in terms of mean of ventilated days. We use a default value of 40% for p_t. We also run a sensitivity analysis on this proportion ranging from 25% to 100%.

(iiii) Temporal Correlation:

The DLM forecasting model yields an estimate of a one-week lag temporal correlation in ILI hospital admissions. We detail these estimates in separate work [24], and summarize the results in Table 2.2. These estimated temporal correlations vary by week and by region. In half the regions, the peak-week correlation is the minimum week-to-week correlation (with the preceding week), during the April-December 2009 period. In the other half of the regions, the minimum week-to-week correlation occurs between the week prior to the peak week and its preceding week.

Table 2.2: Temporal correlation in the DLM forecasting model between consecutive weeks, April-December 2009. When the peak-week correlation is not the minimum correlation over the nine months, the minimum instead occurs the week before the peak week. Source: DSHS [24].

Region	Minimum	Peak-week	Median	Maximum
HSR 1	0.38	0.38	0.44	0.46
HSR 2/3	0.08	0.11	0.28	0.28
HSR 4/5N	0.19	0.19	0.23	0.24
HSR 6/5S	0.32	0.34	0.64	0.65
HSR 7	0.19	0.20	0.33	0.35
HSR 8	0.16	0.16	0.29	0.30
HSR 9/10	0.08	0.12	0.42	0.43
HSR 11	0.07	0.07	0.20	0.21

Table 2.3 lists estimates of regional peak-week demands for ventilators for the mild scenario based on the April-December 2009 hospital discharge data and the other parameters we describe above. We see that the means range from 8.59 to 66.83 while all the coefficients of variation (CVs) (i.e., standard deviation divided by the mean) are below 0.4. In Section 2.5 we also perform a sensitivity analysis on CV by scaling those values in the mild scenario.

Table 2.3: Estimated regional peak-week demands for ventilators in the mild scenario. These estimates are based on April-December 2009 hospital discharge data in Texas. All the regional peak demands have a coefficient of variation below 0.40 although the means range from 8.59 to 66.83.

Region	Mean (units)	Standard deviation (units)	Coefficient of variation
HSR 1	8.59	3.09	0.36
HSR 2/3	66.83	11.31	0.17
HSR 4/5N	12.93	3.48	0.27
HSR 6/5S	40.20	7.79	0.19
HSR 7	25.14	6.01	0.24
HSR 8	22.41	5.29	0.24
HSR 9/10	17.55	4.66	0.27
HSR 11	35.97	7.70	0.21

2.4.2 Scaling Ventilator Demand for the Moderate and Severe Scenarios

The mild scenario that we consider in this chapter is based on the distributions of peak hospitalizations estimated by applying the DLM forecasting model to 2009 H1N1 hospital discharge data for Texas [23]. Since we do not have access to comparable hospitalization data for 1958/68 (moderate) and 1918 (severe) influenza pandemics, we scale our 2009 DLM forecasts in order to achieve similar forecasts for moderate and severe pandemics. The HHS [65] uses the values reported in Table 2.4 for planning purposes. The HSC [68] uses a similar table with identical values

for illness, outpatient care, and deaths, and values of hospitalization, ICU care, and mechanical ventilation that are lower by about 17% for the moderate scenario and 14% for the severe scenario. The CDC's median estimate of hospitalizations for the 2009 H1N1 pandemic (April 2009 - April 2010) is 275,000. Using Table 2.4, we scale our 2009 forecasts by $865/275 = 3.14$ to yield a moderate pandemic scenario and by $9,900/275 = 36$ to yield a severe scenario. The scaling preserves the temporal correlations that the data suggest for the 2009 mild scenario.

Table 2.4: Number of illnesses, healthcare utilization, and deaths associated with moderate and severe pandemic influenza scenarios. Source: HHS [65].

Characteristic	Moderate (1958/68-like)	Severe (1918-like)
Illness	90 million (30%)	90 million (30%)
Outpatient medical care	45 million (50%)	45 million (50%)
Hospitalization	865,000	9,900,000
ICU care	128,750	1,485,000
Mechanical ventilation	64,875	742,500
Deaths	209,000	1,903,000

One caveat associated with our method of scaling the 2009 mild scenario to achieve moderate and severe scenarios concerns the shape of the epidemic curve. The HSC planning scenarios [68] have 20-22% of total pandemic hospitalizations occurring in the peak week along with 34-37% of the demand for ICU beds and 34-37% of the demand for ventilated beds occurring in the peak week. The HSC indicates that this comes from calibrating the shape of the epidemic curve to data from the 1918 pandemic, where that data suggest a factor of seven difference in the mortality rate from the beginning of the pandemic to its peak. We do not attempt such a calibration and instead use the shape of the 2009 H1N1 pandemic scenario to guide our moderate and severe pandemics. As a result, for all three scenarios, the mean demand for ventilated beds in our peak weeks, summed across all eight HSRs, accounts for 6%

of total mean demand during the pandemic period. Our peak-week forecasts indicate Texas would account for 4.5-4.7% of the national demand for mechanical ventilation according to the HSC peak-week scenarios for moderate and severe pandemics.

2.5 Results and Discussion

For each of the three pandemic scenarios, we calculate the sizes of the stockpiles needed to ensure that ventilator demand is satisfied with a specified level of risk. We quantify risk in two different ways: expected unmet demand (the expected number of patients not receiving required ventilation, summed across all HSRs) and probability of unmet demand (the probability that at least one patient in the state will not receive required ventilation).

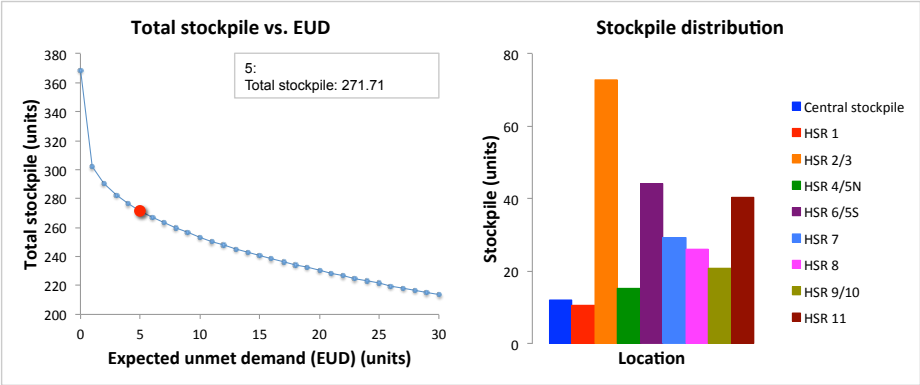
2.5.1 Results for Mild, Moderate, and Severe Pandemics

Figure 2.6 shows the results for the mild pandemic scenario. In this case, we solve the problem for both the central and regional stockpiles. The left-hand curve in Figure 2.6(a) shows the tradeoff between the expected unmet ventilator demand on the x-axis and the total stockpile of ventilators on the y-axis. We highlight a point with a bold dot on the tradeoff curve, indicating that a stockpile of about 272 ventilators is needed to ensure an EUD of at most five ventilators. The bar chart on the right shows the breakdown of the 272 ventilator stockpile into central and regional stockpiles. Figure 2.6(b) is similar except that the x-axis depicts the probability that there is unmet demand for ventilators. A stockpile of 272 ventilators corresponds to a 30% probability of unmet demand somewhere in the state-wide system. The event of unmet demand means that at least one HSR has a shortfall of ventilators when combining their regional stockpile with that received from the central

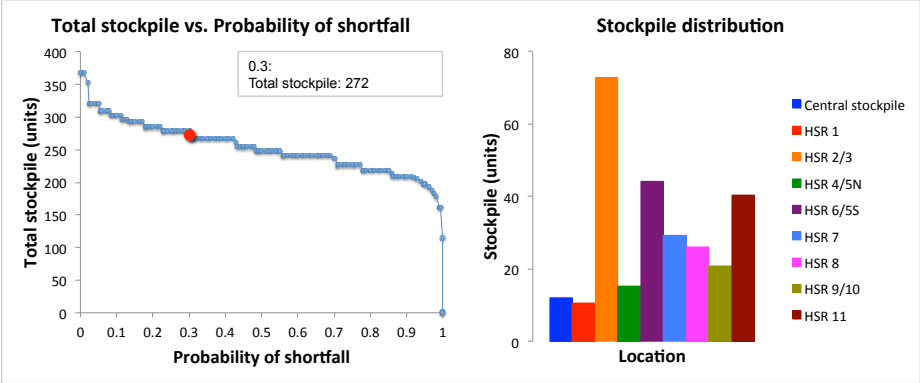
stockpile. We emphasize that the optimization model only minimizes the number of ventilators required to achieve various levels of EUD. The corresponding probability of unmet demand is computed after solving this optimization model. While the probability of unmet demand typically decreases in a stepwise fashion as the size of the stockpile increases (see Figure 2.6(b)), it is possible for the probability to increase as the stockpile increases. This atypical behavior can occur because a low probability-major shortfall event can significantly increase EUD but not the probability of unmet demand, while a high probability-small shortfall event has the reverse effect.

These results are based on 1,000 i.i.d. observations drawn from the multivariate normal distribution representing peak demand for ventilators. The tradeoff in Figure 2.6 indicates that as we require a stockpile that ensures a smaller value of EUD, the magnitude of the requisite stockpile grows sharply. On the other hand, once the available stockpile drops to a sufficiently low level, the tradeoff between total stockpile and the expected value of unmet demand is essentially linear. The mix of ventilators stockpiled centrally versus regionally is biased strongly in favor of stockpiling in the regions. We investigate the sensitivity of this result to the wastage factor and the region-to-region correlation later in this section.

Figure 2.7 depicts the results of solving for both the central and regional stockpiles under the moderate and severe pandemic scenarios. As we describe in Section 2.4, we scale the mild scenarios by factors of 3.14 and 36 to achieve distributions for moderate and severe hospitalizations, and we assume that 25% of hospitalizations require ICU care under the moderate and severe scenarios (compared to 20% for the mild scenario). The optimal stockpiling solutions scale directly as $(0.25/0.2) \cdot 3.14 = 3.93$ and $(0.25/0.20) \cdot 36 = 45$. As a result, total stockpiles of



(a) Expected unmet demand versus number of ventilators stockpiled



(b) Probability of shortfall versus number of ventilators stockpiled

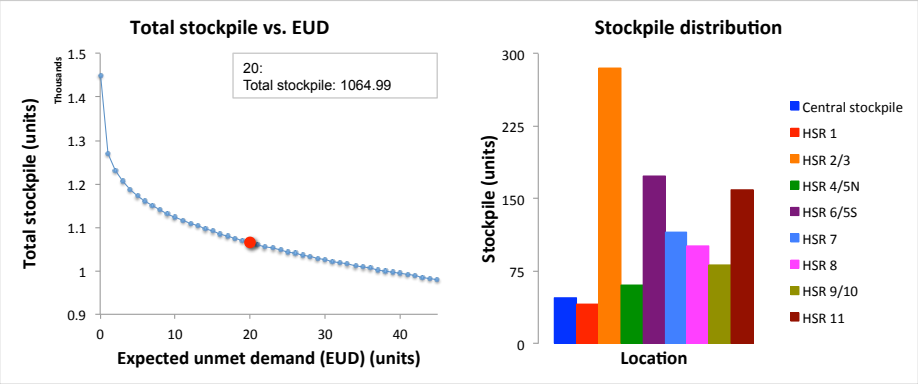
Figure 2.6: Stockpiling results for the mild pandemic scenario. We quantify the risk associated with ventilator stockpiles in terms of both (a) expected number of ILI patients not receiving necessary ventilation and (b) probability that at least one ILI patient in the state will not receive necessary ventilation. Solving the optimization model yields the stockpiles necessary to ensure a maximum level of expected unmet demand, and the probability of a shortfall is calculated after the fact. The left-hand side of Figure 2.6(a) shows the total stockpile (summed across the eight HSRs and the central stockpile) versus the magnitude of the expected unmet demand. An expected unmet demand of five ventilators corresponds to a total stockpile of about 272 ventilators (shown by the larger blue circle on the curve and the values in the box at top right corner of the graph). The bar chart on the right-hand side of Figure 2.6(a) depicts the associated portfolio of centrally stockpiled and regionally stockpiled ventilators. Figure 2.6(b) is similar except that the x-axis shows the probability that there is unmet demand in at least one HSR. A stockpile of 272 ventilators corresponds to a probability of unmet demand of 30%.

$1,065 \approx 3.93 \cdot 272$ and $12,204 \approx 45 \cdot 272$ correspond to an EUD of $20 \approx 3.93 \cdot 5$ and $225 \approx 45 \cdot 5$ under these respective scenarios, as shown in Figure 2.7.

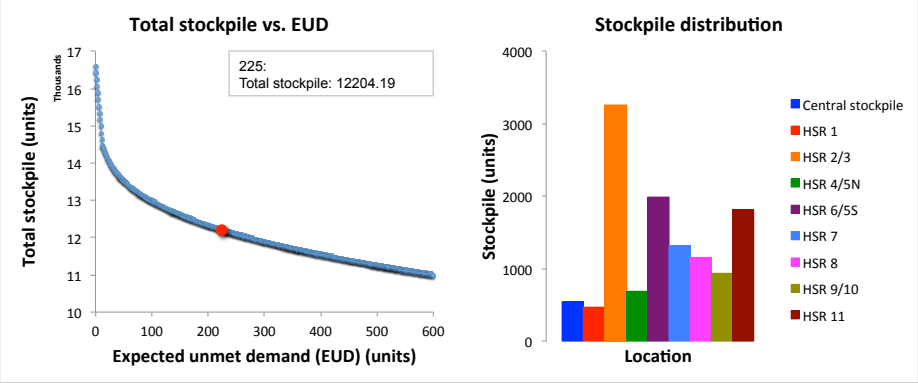
We also describe an example in which we solve only for the size of the central ventilator stockpile. In this case, we assume that the regional stockpiles are at the levels described in DSHS reports [57], shown in Table 2.5. Under the moderate pandemic scenario, these regional stockpiles are several standard deviations above the forecasted mean demands and sufficient for all regional demands in all scenarios sampled. Therefore, the solution suggests that no central stockpile is needed. Under the severe scenario, however, these stockpiles are inadequate. The regions hold a total of 3,730 ventilators. The sum of the mean demands across the eight HSRs under the severe scenario is forecasted to be 10,333 ventilators. Given the probabilistic nature of the demand and optimistically assuming zero wastage, the model indicates that a central stockpile of 6,763 ventilators will be necessary to limit the probability of unmet demand to 50%, with an expected shortfall of 737 ventilators. (As of 2006, the SNS maintained 5,000-6,000 ventilators for distribution to all of the states in the U.S. [1, 68].)

Table 2.5: Existing regional stockpiles of ventilators in the state of Texas. Source: DSHS [57].

Region	Existing ventilators (units)
HSR 1	151
HSR 2/3	1233
HSR 4/5N	247
HSR 6/5S	742
HSR 7	247
HSR 8	458
HSR 9/10	287
HSR 11	365



(a) Moderate pandemic scenario



(b) Severe pandemic scenario

Figure 2.7: Stockpiling results for the moderate and severe pandemic scenarios. Part (a) shows the tradeoff between expected number of ILI patients not receiving necessary ventilation across all HSRs for the moderate pandemic scenario and part (b) shows that for the severe pandemic scenario. The bar charts on the right-hand side depict the associated portfolio of centrally stockpiled and regionally stockpiled ventilators for the moderate and severe scenarios, respectively. The optimal stockpiles for the (a) moderate and (b) severe scenarios scale with $(0.25/0.20) \cdot 3.14 = 3.93$ and $(0.25/0.20) \cdot 36 = 45$ over the mild scenario of Figure 2.6.

2.5.2 Sensitivity Analysis

In addition to analyzing three different pandemic scenarios, we investigate the effects of changing key model input parameters on the recommended stockpiling strategies. We first consider the proportion of hospitalizations requiring ICU care, the proportion requiring mechanical ventilation, and the proportion requiring two weeks of ventilation. Then, we consider the wastage proportion and the region-to-region correlation coefficient, as well as the CVs of regional peak-week demands for ventilators.

2.5.2.1 ICU, Ventilation, and Two-Week Proportions

Changing the proportion of hospitalizations requiring ICU care and/or the proportion requiring mechanical ventilation directly scales our demand for ventilators, similar to scaling the mild scenario to achieve moderate and severe scenarios. As a result, if we increase the proportion of ICU patients requiring ventilation from 0.5 to 0.67 then there is no need to resolve the model. The optimal stockpiles will scale by a factor of $0.67/0.50 = 1.34$. The same result holds if we change the ICU proportion or simultaneously change both the ICU and ventilator proportions.

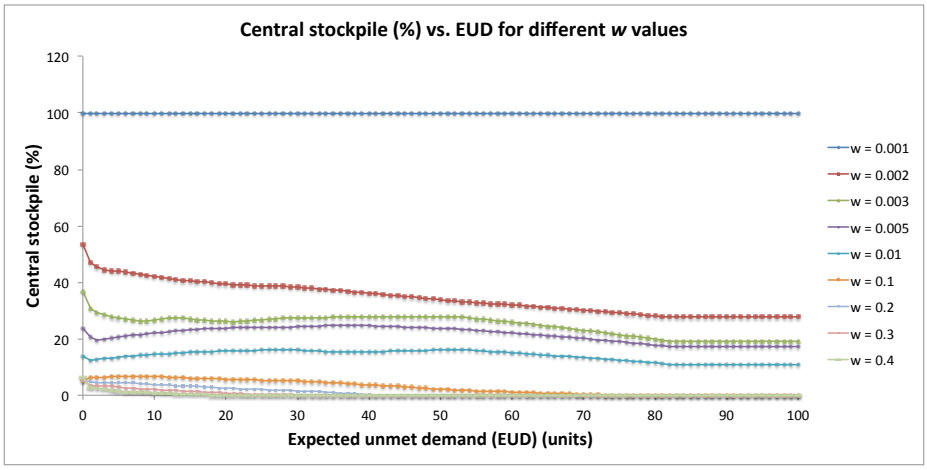
The effect of changing the proportion of ventilated patients requiring mechanical ventilation for two weeks is more subtle. For simplicity, suppose we were to increase the proportion from 0 to 1. Then, the mean demand would not simply double because consecutive weeks do not have identical means. For the 2009 hospitalization data in Texas, increasing this proportion from 0 to 1 increases the total mean demand, summed across all regions, by a factor of 1.96. This simplistic scaling rule suggests that the stockpiles should increase by a factor of $2/1.4 = 1.43$ when

increasing the two-week proportion from 0.4 to 1, or by a factor of 1.42 when examining the increase in total mean demand for the 2009 hospitalization data. The actual increase in the optimized stockpile depends on the risk level we set for the EUD with the increase ranging from a factor of 1.38 when the limit on EUD is near zero, up to 1.42-1.43 for limits on EUD less than five for the mild scenario. The actual increase is smaller than the factor of 1.42-1.43 when EUD is near zero because the spread of the distribution (i.e., the coefficient of variation) decreases as the two-week proportion grows.

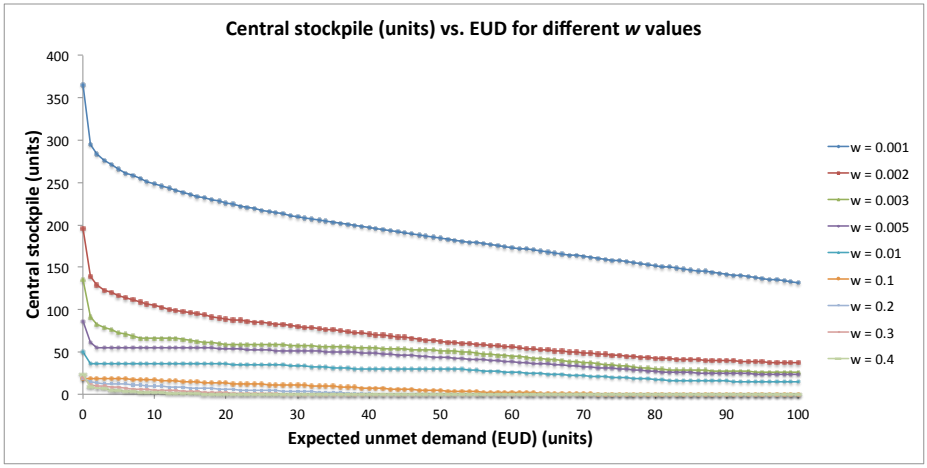
2.5.2.2 Wastage Proportion and Region-to-Region Correlation

The default value of wastage is $w = 0.2$, or 20%, meaning that one in five ventilators distributed to a region is not used effectively if needed. The size of the central stockpile is a small fraction of the total stockpile under the default settings of other parameters. For example, for the mild scenario at an EUD of five ventilators, the central stockpile is just 4.4% of the total stockpile. As the wastage parameter shrinks to zero, we expect the percentage of the total stockpile held centrally to increase. Table 2.6 shows that it indeed increases, but that wastage has to drop to surprisingly small values for the central stockpile to grow significantly. The table also shows that as the region-to-region correlation shrinks from its default value of 0.70 to 0.55 the centrally held stockpile becomes more sensitive to the wastage parameter.

Figure 2.8 depicts the central stockpile versus EUD for the same values of the wastage parameter (w) as shown in Table 2.6. Figure 2.9 is an analogous figure except that we depict the central stockpile for various values of ρ_{HSR}, the region-to-region correlation coefficient.

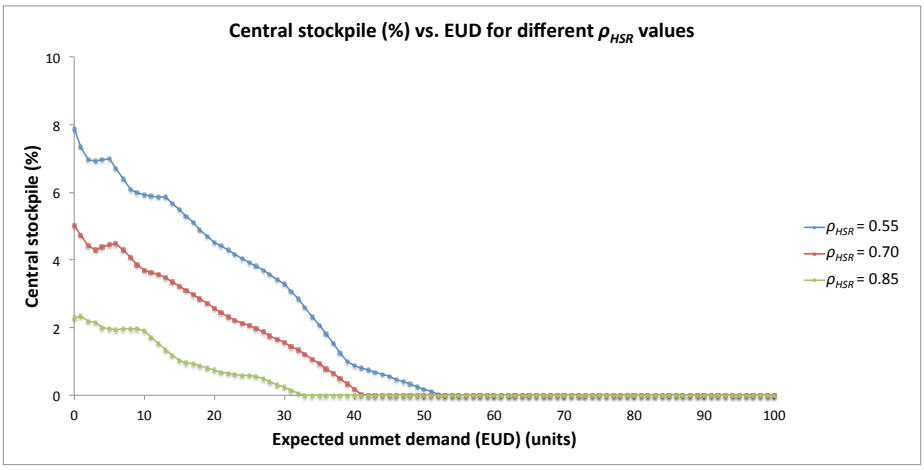


(a) Central stockpile as a percentage of the total stockpile

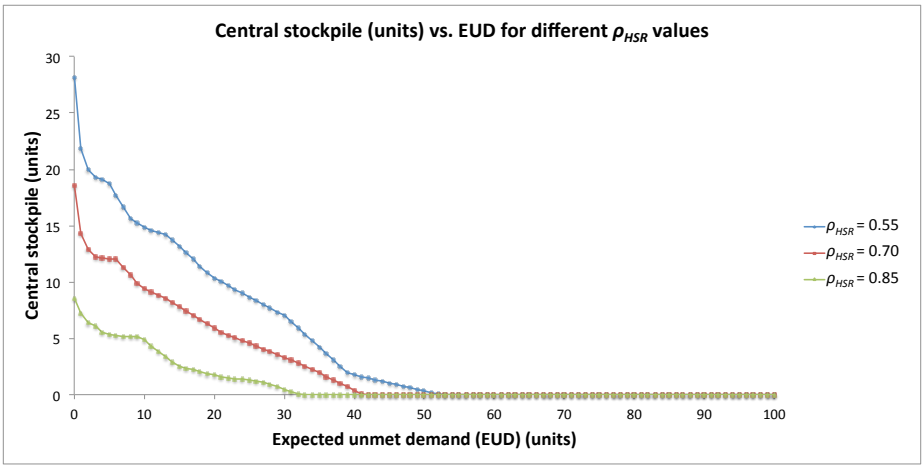


(b) Central stockpile in units of ventilators

Figure 2.8: Central stockpile versus expected unmet demand (EUD) for various w values. The baseline result, i.e., the mild pandemic scenario, corresponds to $w = 0.2$, or 20 %. Part (a) shows the change in the percentage of the stockpile held centrally with the growth of EUD while part (b) shows the change in the number of ventilators held in the central stockpile.



(a) Central stockpile as a percentage of the total stockpile



(b) Central stockpile in units of ventilators

Figure 2.9: Central stockpile versus expected unmet demand (EUD) for various ρ_{HSR} values. The baseline result, i.e., the mild pandemic scenario, corresponds to $\rho_{HSR} = 0.70$. Part (a) shows the change in the percentage of stockpile held centrally with the growth of EUD and part (b) shows the change in the number of ventilators held in the central stockpile.

Table 2.6: Percent of stockpile held centrally to achieve expected unmet demand of at most five ventilators in the mild scenario, for various combinations of the wastage parameter (w) and region-to-region correlation in peak ventilator demand (ρ_{HSR}). The baseline result of 4.4% is indicated in bold.

w (%)	Central Stockpile (%)		
	$\rho_{HSR} = 0.55$	$\rho_{HSR} = 0.70$	$\rho_{HSR} = 0.85$
40	2.8	1.5	0.2
30	4.7	2.9	0.9
20	7.0	**4.4**	2.0
10	9.8	6.7	3.5
1	18.0	13.5	10.1
0.5	25.4	20.6	17.0
0.3	32.8	27.4	22.7
0.2	48.2	43.9	39.3
0.1	100	100	100

2.5.2.3 Coefficient of Variation (CV)

We scale the CVs of the mild scenario, as shown in Table 2.3, by factors of 0.5 to 3 to investigate the effect of the variability of the demand distribution on the recommended strategies for stockpiling. Table 2.7 shows the number of ventilators centrally stockpiled under different scaling values. For each value for the coefficient of variation, we see that the central stockpile tends to decrease (in percentage and absolute terms) as EUD grows. In models (2.1) and (2.2), central stockpiling has the advantage of being able to distribute ventilators to regions after the realization of regional demands. Hence, we might expect an increase in central stockpiling when there is greater uncertainty in regional peak-week demands. Table 2.7 shows that when doubling and tripling the CVs of the mild scenario, the central percentage grows, both in percentage and absolute terms. Figure 2.10 shows further results in the spirit of Table 2.7.

Moreover, when doubling the CVs of the mild scenario, we see the central

stockpile goes up slightly along with EUD, when EUD is small, it goes down after EUD is greater than 20. The same tendency exists when tripling the CVs of the mild scenario. There are two reasons for it. First, when we smooth the optimal solutions between nearby L values, we smooth the number of ventilators stockpiled by DSHS and regions, instead of their percentage. The number of centrally stockpiled ventilators does go down monotonically for all variation levels, as shown in Table 2.7(b). The other reason is the proportion of wastage when distributing the central stockpile to regions. When constraining EUD to be close to zero, we should satisfy nearly all regional demands under any realized scenario. In this case, we may want to stockpile more in regions (in terms of percentage of total stockpile) to avoid the wastage.

Our results indicate that under the baseline scenario the percentage of the stockpile of ventilators held centrally is small. There are three parameters that can lead to changes in this result. If the percentage of ventilators wasted, when shipped from the central stockpile to the regions, is small, then the central stockpile grows. If the region-to-region correlation coefficient shrinks, then the central stockpile grows. And, as the coefficient of variation in demand grows, the central stockpile grows. That said, the values of these parameters must change significantly in order for the central stockpile to grow to (say) 10% of the total stockpile.

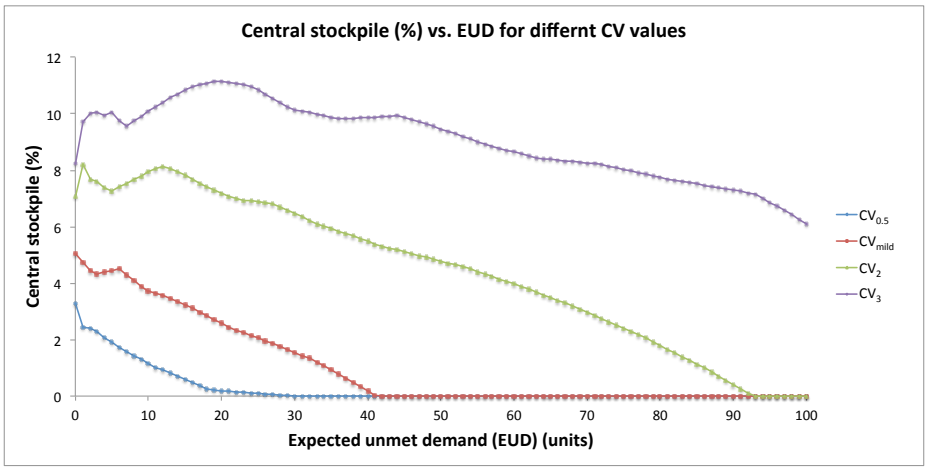
Table 2.7: Stockpile held centrally to achieve various expected unmet demand for different levels of CV. We scale the CVs in the mild scenario by factors of 0.5 to 3. The subscriptions of CV indicate the scaling factor. Part (a) shows the central stockpile in terms the percentage of total stockpile and part (b) show the recommended ventilators stockpiled centrally. We see there is a tendency to reduce the amount of centrally held ventilators when the allowable expected unmet demand is larger, as well as when the variation level of regional peak demands is smaller.

(a) Central stockpile in percentage (%)

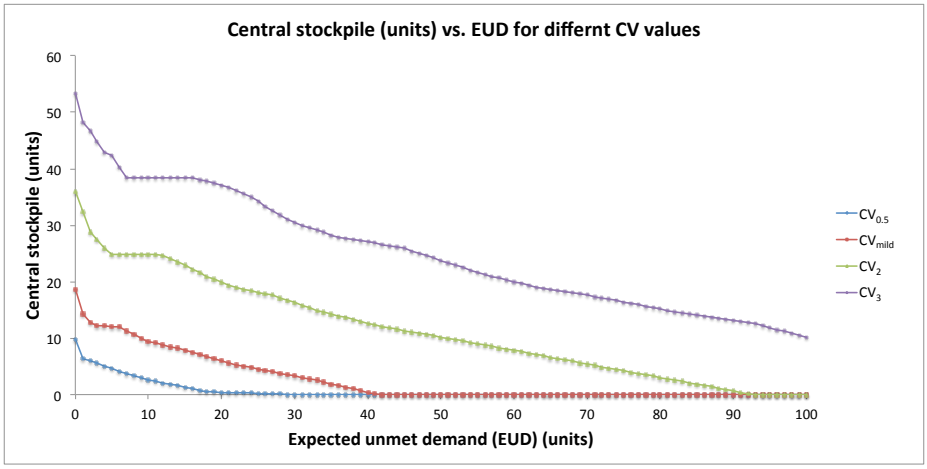
Expected unmet demand (units)	Central Stockpile (%)			
	$CV_{0.5}$	CV_{mild}	CV_2	CV_3
5	1.9	4.4	7.2	10.0
10	1.1	3.7	7.9	10.1
20	0	2.6	7.2	11.2
30	0	1.6	6.5	10.1
50	0	0	4.8	9.5
100	0	0	0	6.1

(b) Central stockpile in units

Expected unmet demand (units)	Central Stockpile (units)			
	$CV_{0.5}$	CV_{mild}	CV_2	CV_3
5	5	12	25	42
10	3	9	25	38
20	0	6	20	37
30	0	3	16	30
50	0	0	10	24
100	0	0	0	10



(a) Central stockpile as a percentage of the total stockpile



(b) Central stockpile in units of ventilators

Figure 2.10: Central stockpile versus expected unmet demand (EUD) for various CV values of peak demand for ventilators. We scale the CVs in the mild scenario by factors of 0.5 to 3. The subscriptions of CV indicate the scaling factor. The baseline result, i.e., the mild pandemic scenario, corresponds to $\mathrm{CV}_{\mathrm{mild}}$. Part (a) shows the change in the percentage of the stockpile held centrally with the growth of EUD while part (b) shows the change in the number of ventilators held in the central stockpile.

Chapter 3

Optimizing Allocation of Pandemic Influenza Vaccines

3.1 Introduction

The effectiveness of vaccination to thwart the spread of influenza in a pandemic has received significant attention in the literature [43, 47]. Influenza strains vary over time, and the process of manufacturing vaccines is complex. Hence, demand can exceed vaccine supply, even for seasonal influenza and particularly during a pandemic [64]. We describe a framework for allocating and distributing vaccines of different types during an influenza pandemic.

When vaccine supply is limited, a natural question is how to best allocate available vaccines. Medlock and Galvani [35] use an age-structured transmission model to optimally allocate vaccines in the U.S. for pandemic influenza according to five criteria: deaths, infections, years of life lost, contingent valuation, and economic costs. By using data from the 1918 and 1957 pandemics, they recommend prioritizing schoolchildren and adults aged 30-39 years because schoolchildren are most responsible for transmission, and their infectious parents later spread the disease further. Medlock and Galvani [35] also show that distributing vaccines uniformly to people aged 5-19 years has a similar performance as their optimal allocation for all five criteria. Keeling and White [26] use three simple extended susceptible-infectious-recovered models to address whether targeting risk groups, age groups, or spatial regions for

vaccination could reduce the predicted number of influenza cases in Great Britain. Based on 2009 H1N1 data, they show that prioritizing high-risk groups (rather than high-transmission groups), age groups of 5-14 years and then 15-24 years, or regions of the country most affected generally leads to a greater reduction in the number of cases. Keeling and White [26] further indicate that these three priorities may vary as the epidemic progresses.

It can be difficult to observe, and react to, the geographic spread of influenza and certain parameters that drive disease spread models. For this reason, we may seek a "fair" allocation. Wu et al. [70] use a simulation model to assess the performance of different distributions of pre-pandemic influenza vaccines in the U.S. Compared to other discretionary policies, a pro-rata policy may not be the most effective in terms of the number of infections averted. However, the performance of discretionary polices are sensitive to parameters that are difficult to know in advance. Wu et al. [70] conclude that the pro-rata policy is simple (no need for epidemiological information), robust (compatible performance under various scenarios), and equitable (equal chance of vaccination), and hence the pro-rata policy may be a sensible approach. Araz et al. [3] argue that vaccine distribution should be differentiated according to the demographic and spatial structures of communities because the activity and severity of the 2009 H1N1 influenza varied considerably among age groups and locations. By examining 2009 H1N1 data from counties in the state of Arizona, they conclude that prioritizing counties which are expected to experience the latest epidemic waves reduces overall attack rate most effectively. However, Araz et al. [3] also predict that the pro-rata policy will be an effective strategy when considering both the attack rate and the waiting period for those seeking vaccines. Matrajt et al. [34] use an infection transmission model coupled with a genetic algorithm to dynamically distribute vac-

cines to two age-groups: children and adults. In a network of 16 cities in Southeast Asia, they find that their allocation outperforms the strategy of allocating vaccines proportional to population, in terms of illness attack rate, given that vaccination occurs within the first weeks after a pandemic starts. That said, they acknowledge the potential difficulties of an uneven geographic distribution from the view of ethics and fairness and suggest that vaccinating only children in each city in proportion to the children's population can be an effective solution.

During the 2009 H1N1 pandemic, when delivering donated and purchased vaccine doses to qualified countries, the World Health Organization (WHO) tried to ensure equitable access to vaccines among the countries. The requests from these recipient countries were prioritized based upon epidemiological, programmatic, and other criteria [44]. At the same time, the U.S. Centers for Disease Control and Prevention (CDC) allocated H1N1 vaccines to states of the U.S. in proportion to their populations, seeking equitable vaccination coverage across the states [53]. Guidance on pandemic vaccination from the U.S. Department of Health and Human Services and the U.S. Department of Homeland Security [67] states explicitly that vaccines will be allocated to the states in proportion to each state's population. And, in order to ensure fairness and uniformity across the U.S. they strongly suggest that each state should follow national guidance. Equitable *distribution* of vaccines may be desirable, but disparities in influenza immunization *uptake* exists for several other reasons. According to Logan [32], there are a number of factors that account for disparities in vaccination rates in the U.S. These include economic factors (including insurance status), perceptions of the health risks of both influenza and vaccines, and a lack of trust of healthcare systems. Logan [32] further proposes strategies to mitigate the effect of these factors.

Fairly allocating limited resources to users has received significant attention in communication networks where we assign rates to users of a network that has link capacity constraints. The objective is to maximize total throughput while ensuring a certain type of fairness. There are several types of fairness proposed in the communication networks literature, including max-min fairness, proportional fairness, and (p, α)-proportional fairness [30, 61]. Max-min fairness [8] has been a well-known and widely-adopted criterion for egalitarian allocation. When assigning rates to users, max-min fairness prioritizes users with small demands and evenly distributes remaining rates to other users with high demands (see [30] for definition and a concrete example). In the literature, there is a class of so-called water-filling algorithms used to obtain max-min fair solutions [12]. Instead of just considering users' demands, Kelly [27] suggests proportional fairness which takes the required resource for a user's demand into consideration. The idea of proportional fairness is that each user should have an equal share of resource use so that if two users have the same demand, the user who requires more resource (i.e., capacity on links) use will get less rate assigned (again, see [30] for definition and a concrete example). Note that this notion of proportional fairness differs from our notion of a proportionally fair coverage when allocating vaccines, which we define in Section 3.2. Mo and Walrand [37] generalize Kelly's proportional fairness and propose (p, α)-proportional fairness. Max-min fairness is a limiting case in which $p = (1, \ldots, 1)$ and $\alpha \to \infty$, and proportional fairness is a special case in which $p = (1, \ldots, 1)$ and $\alpha = 1$ [37, 61].

Researchers have formulated various communication network problems to achieve these types of fairness (e.g., [27, 37, 61]). In particular, Boyd and Vandenberghe [9] (on page 245) formulate an optimization problem that can be interpreted as allocating limited power to a set of communication channels which has various levels of

power already allocated. They maximize a sum of logarithmic functions to obtain an allocation where each channel has no less than a fair level of power assigned in total. Coluccia et al. [12] discuss another basic resource allocation problem where a single type of limited resource must be shared among a set of users with heterogeneous demands. They formulate it as a minimization problem and prove a sufficient condition for the objective function to achieve max-min fairness. Least squares or maximum entropy are two examples satisfying such sufficient condition. As we mention above, an optimal max-min fair allocation assigns the resource to fulfill users with small demands and bring all other users with high demands to the same level. Coluccia et al. [12] also prove that max-min fairness is equivalent to proportional fairness in the problem they consider.

The models and tools we describe above from the literature do not apply in our setting of seeking a fair allocation of vaccines for two reasons. We have multiple types of vaccines and we partition "users" into several target populations. Only a subset of the target populations are eligible to receive each vaccine type, and we can assign different weights to each target population. This chapter is motivated by a project with the Texas Department of State Health Services (DSHS), and the framework we describe is designed to achieve proportionally fair coverage of pre-specified priority groups across the 254 counties in the state of Texas, as informed by weights on those priority groups and/or counties. We use the weights to address different desired coverage among priority groups and/or counties. We use coverage here to mean the proportion of a population that has access to vaccines, rather than the actual vaccination rate. We acknowledge that equal accessibility does not mean equal vaccination rate, but unequal accessibility is one primary cause of vaccination disparities.

In 2009, after receiving vaccine doses from the Strategic National Stockpile (SNS) via the CDC, the state of Texas allocated vaccines to the state's Registered Providers (RPs), Local Health Departments (LHDs), and Health Service Regions (HSRs). The former two allocations were driven by requests from RPs and LHDs. The vaccine doses going to HSRs were allocated at DSHS's discretion, in contrast to the request-based system for RPs and LHDs. In supply chain networks, a system is often classified as push distribution or pull distribution [21, 49]. In a pull-based system, the distribution decisions are driven by customer requests while in a push-based system, they are based on the supplier's forecasts. Thus, we can categorize RP and LHD allocations as a pull-based system and HSR allocation as a push-based system. Throughout the 2009 H1N1, DSHS used the first two channels to facilitate distributing vaccines to places in need and used HSR channel to boost vaccination coverage in counties where an insufficient number of doses were distributed to RPs. Moreover, during an influenza pandemic vaccines are typically manufactured as the pandemic unfolds, and states receive vaccines from the SNS on a weekly basis. As we describe in further detail below, our modeling framework takes as input all doses allocated to date to RPs, LHDs, and HSRs, and recommends as output HSR allocations targeted to priority groups by vaccine type, all at the geographic resolution of counties.

The organization of the remainder of this chapter is as follows. In Section 3.2, we detail the modeling framework that takes as input priority groups, vaccine types and available doses, as well as coverage to date, and gives as output vaccine allocation targeted to priority groups by vaccine type at the geographic resolution of counties to maximize proportional fairness, as informed by user-specified weights. In Section 3.3, we describe the data used to simulate vaccine allocation during the 2009 H1N1 pandemic. In Section 3.4, we present simulation results and perform a sensitivity analysis

on the portion of total doses reserved by DSHS for allocating to HSRs. In Section 3.5, we discuss the results and potential use of the modeling framework.

3.2 Modeling Framework

We build two optimization models and describe a post-processing step which together allocate available HSR doses to counties and, in turn, to pre-specified priority groups, seeking proportional fairness with policy simplicity and regional equity in mind. First, we build an optimization model that seeks a proportionally fair coverage for each county-priority group pair. The inputs include (i) coverage to date, weight, and population for each county-priority group pair, (ii) available HSR doses, and (iii) vaccine suitability rules. The model then provides as output the optimal coverage rate for each county-priority group pair. Next, the second optimization model takes as input (i) the optimal coverage rates of the first model, (ii) a sub-optimality tolerance for these coverage rates, and (iii) vaccine suitability rules. The second model aims to maximize policy simplicity and regional equity while ensuring the gap between the optimal coverage and the resulting final coverage is within the pre-defined tolerance. Lastly, we perform a post-processing step on the allocation from the second model to obtain integral allocations. The flow chart of the modeling framework is shown in Figure 3.1, and the details for each of these steps are described in the remainder of this section.

3.2.1 Model Assumptions

The models we formulate, and associated analyses we carry out, in this chapter are based on the following assumptions:

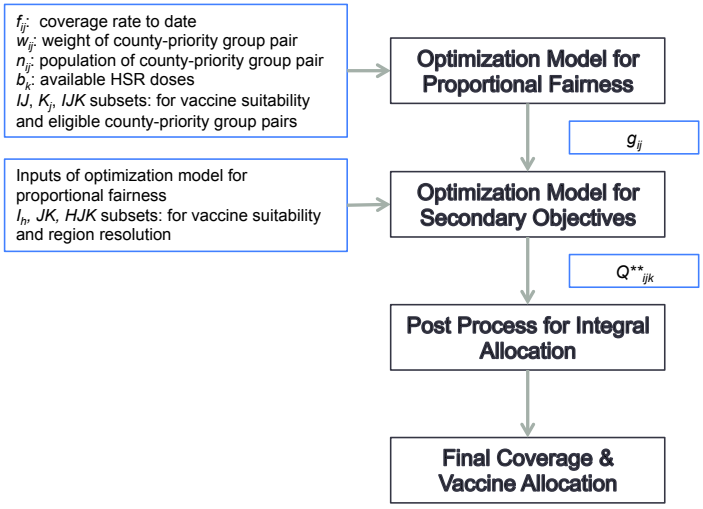


Figure 3.1: The steps and the associated inputs and outputs of the modeling framework. The first optimization model seeks proportionally fair coverage and the second one accounts for secondary objectives: policy simplicity and regional equity, while ensuring near optimality for proportional fairness. The post-processing step makes sure no fractional doses are allocated and then outputs the resulting final coverage and allocation.

1. Vaccines are allocated to counties and, in turn, to priority groups via RPs, LHDs, and HSRs. The former two allocations are driven by requests from RPs and LHDs, which are given to our models as input. The doses DSHS reserves for allocation to HSRs are allocated to counties and, in turn, to priority groups at DSHS's discretion.

2. The state of Texas is partitioned into eight health service regions, as shown in Figure 2.1, and the 254 counties are divided into eight groups as a result.

3. A county in Texas is either served by an LHD or an HSR, but not both. Out of the 254 counties in Texas, there are 65 counties served by LHDs and the other 189 by HSRs. As a result, only these 189 counties are eligible for discretionary HSR doses.

4. A person will seek vaccination only within his/her county of residence and can be vaccinated by one dose of any suitable vaccine type.

5. The suitability of vaccine type-priority group pairs is the same in each county.

3.2.2 Model Notation

We use the following notation in this chapter.

Indices and Sets

$i \in I$: counties

$j \in J$: priority groups

$k \in K$: vaccine types

$h \in H$: health service regions

I_h = $\{i \in I$: counties that are in health service region $h\}$

K_j = $\{k \in K$: vaccine types for which priority group j is eligible$\}$

IJ = $\{(i,j) \in I \times J : n_{ij} > 0$ and $f_{ij} < 1\}$ (see below for definitions of n_{ij} and f_{ij})

JK = $\{(j,k) \in J \times K : k \in K_j\}$

IJK = $\{(i,j,k) \in I \times J \times K : (i,j) \in IJ$ and $k \in K_j\}$

HJK = $\{(h,j,k) \in H \times J \times K : h \in H,\ j \in J$ and $k \in K_j\}$

Data and Parameters

m_{ijk} : doses of vaccine type k already allocated to priority group j in county i

n_{ij} : population of priority group j in county i

w_{ij} : weight of priority group j in county i

b_k : available HSR doses of vaccine type k

f_{ij} : coverage rate of priority group j in county i from doses already allocated

ϵ : tolerance for optimal proportional fairness when considering secondary objectives

M : large value for big M method

Decision Variables

Q_{ijk} : number of vaccines of type k allocated to priority group j in county i

V_{jk} : 1 if $\sum_{i \in I:(i,j,k) \in IJK} Q_{ijk} > 0$; 0 otherwise

Y_{hjk} : regional coverage of priority group j in region h from available doses of type k

$\overline{Y}_{jk}$: average regional coverage of priority group j from available doses of type k

3.2.3 Optimization Model for Proportional Fairness

We build a model that optimally allocates available HSR vaccine doses for different types of vaccines to county-priority group pairs. The new allocation accounts for previous allocations from RPs and LHDs, and for any allocations from HRSs in previous weeks. By an optimal allocation, we mean that we seek to bring all under-served county-priority group pairs to a proportionally fair level, using available HSR doses; i.e., we seek proportional fairness, as informed by user-specified weights on each county-priority group pair.

The optimization model for finding a proportionally fair allocation is as follows:

$$\min_{Q} \quad \sum_{(i,j)\in IJ} w_{ij} n_{ij} \left[1 - \frac{1}{w_{ij}} \left(f_{ij} + \frac{\sum_{k\in K_j} Q_{ijk}}{n_{ij}} \right)\right]^2 \tag{3.1a}$$

$$\text{s.t.} \quad \sum_{(i,j):(i,j,k)\in IJK} Q_{ijk} \leq b_k, \forall k \in K \tag{3.1b}$$

$$f_{ij} + \frac{\sum_{k\in K_j} Q_{ijk}}{n_{ij}} \leq 1, \forall (i,j) \in IJ \tag{3.1c}$$

$$Q_{ijk} \geq 0, \forall (i,j,k) \in IJK. \tag{3.1d}$$

We use f_{ij} to represent the coverage before HSR allocation, which can be calculated as:

$$f_{ij} = \frac{\sum_{k\in K_j} m_{ijk}}{n_{ij}}.$$

An extreme version of an imbalance in previous allocations would be that for a specific county-priority group (i,j) pair, the coverage, f_{ij}, exceeds 1. Given that doses are

scarce, we clearly will allocate no additional doses to that (i, j) pair. Moreover, as state above, we assume only the 189 counties served by HSRs qualify for available HSR doses; i.e., we exclude the other 65 counties from HSR allocation. These kind of exclusions can be easily achieved by defining a subset, IJ, properly. We form five other subsets K_j, JK, IJK, I_h, and HJK. The first three subsets account for suitability constraints governing vaccine type, while the last two subsets are used to balance vaccine allocation among HSRs, which we discuss later in considering secondary objectives.

The inputs of model (3.1) include (i) subsets K_j, IJ, and IJK, (ii) the population, weight, and original coverage of each county-priority group pair, and (iii) available HSR doses of each type. Then, the model provides an optimal allocation, denoted Q^*_{ijk}, and, in turn, the optimal final coverage of each pair, denoted g_{ij}, i.e.,

$$g_{ij} = f_{ij} + \frac{\sum_{k \in K_j} Q^*_{ijk}}{n_{ij}}. \tag{3.2}$$

For simplicity of understanding model (3.1), suppose for the moment that we let the weights be $w_{ij} = 1$ for all county-priority group pairs, (i, j). Then, the objective function in (3.1a) sums the square of the shortage rates,

$$1 - \left(f_{ij} + \frac{\sum_{k \in K_j} Q_{ijk}}{n_{ij}} \right),$$

weighted by the county-priority group population across all such pairs. And, we seek to minimize the sum of these squared population-weighted shortage values.

We add weights ($w_{ij} \geq 1$) for each county-priority group pair to acknowledge their relative importance, where the value of the least important county-priority group pair under consideration is 1. When the weights differ, instead of seeking equal coverage for priority groups in each county, we seek equal coverage when weighted

by the proportionality constants, w_{ij}. So, if one county-priority group has twice the weight of another, we seek twice the coverage rate for the higher priority pair. This can indeed be achieved for two county-priority group pairs under the following conditions: (i) both of their final coverage rates, after HSR doses are allocated, are less than 1, and (ii) these two pairs have at least one common type of available HSR vaccine allocated in an optimal solution.

Constraint (3.1b) ensures that the allocated vaccine doses of type k do not exceed the available HSR doses of that type. Constraint (3.1c) indicates that the maximum coverage is at most 1 for every county-priority group pair receiving available HSR doses. Constraint (3.1d) makes sure that we only allocate a nonnegative number of HSR doses.

Here, we show that this objective function, when combined with the constraints (3.1b)-(3.1d), seeks to provide an allocation of available HSR doses in a manner of proportional fairness.

Theorem 3.2.1. *By using model (3.1) to allocate available HSR vaccine doses, the optimal coverage of two county-priority group pairs, (i, j) and (i', j'), denoted g_{ij} and $g_{i'j'}$, are proportional to their weights, w_{ij} and $w_{i'j'}$, if (i) both g_{ij} and $g_{i'j'}$ as defined in equation (3.2) are less than 1; and, (ii) (i, j) and (i', j') have a common type of available HSR vaccine allocated in an optimal solution.*

Proof. Let λ_k and ν_{ij} be the Lagrangian multipliers of constraints (3.1b) and (3.1c), respectively. Let $F(Q)$ denote the objective function in (3.1a). Then, we have

$$\frac{\partial F}{\partial Q_{ijk}} = 2\left[-1 + \frac{1}{w_{ij}}\left(f_{ij} + \frac{\sum_{k\in K_j} Q_{ijk}}{n_{ij}}\right)\right].$$

As a result, the Karush-Kuhn-Tucker optimality conditions for the convex quadratic program (3.1) (see, e.g., [9]) are then:

Primal feasibility

constraints (3.1b), (3.1c), and (3.1d)

Dual feasibility

$$\lambda_k^* + \frac{\nu_{ij}^*}{n_{ij}} \geq 2\left[1 - \frac{1}{w_{ij}}\left(f_{ij} + \frac{\sum_{k\in K_j} Q_{ijk}^*}{n_{ij}}\right)\right], \forall (i,j,k) \in IJK \tag{3.3a}$$

$$\lambda_k^* \geq 0, \forall k \in K \tag{3.3b}$$

$$\nu_{ij}^* \geq 0, \forall (i,j) \in IJ \tag{3.3c}$$

Complementary slackness

$$\lambda_k^*\left(b_k - \sum_{(i,j):(i,j,k)\in IJK} Q_{ijk}^*\right) = 0, \forall k \in K \tag{3.4a}$$

$$\nu_{ij}^*\left[1 - \left(f_{ij} + \frac{\sum_{k\in K_j} Q_{ijk}^*}{n_{ij}}\right)\right] = 0, \forall (i,j) \in IJ \tag{3.4b}$$

$$Q_{ijk}^*\left\{\lambda_k^* + \frac{\nu_{ij}^*}{n_{ij}} - 2\left[1 - \frac{1}{w_{ij}}\left(f_{ij} + \frac{\sum_{k\in K_j} Q_{ijk}^*}{n_{ij}}\right)\right]\right\} = 0,$$

$$\forall (i,j,k) \in IJK. \tag{3.4c}$$

Consider two (i,j) and (i',j') pairs that satisfy hypotheses (i) and (ii). By hypothesis (i) and complementary slackness condition (3.4b), we have $\nu_{ij}^* = \nu_{i'j'}^* = 0$.

By hypothesis (ii), we have $Q^*_{ijk} > 0$ and $Q^*_{i'j'k} > 0$ for some k and hence by (3.4c) for that k:

$$\frac{2-\lambda^*_k}{2} = \frac{1}{w_{ij}}\left(f_{ij} + \frac{\sum_{k\in K_j} Q^*_{ijk}}{n_{ij}}\right) = \frac{1}{w_{i'j'}}\left(f_{i'j'} + \frac{\sum_{k\in K_j} Q^*_{i'j'k}}{n_{i'j'}}\right).$$

Using the definition of g_{ij} from equation (3.2), we have

$$\frac{g_{ij}}{g_{i'j'}} = \frac{w_{ij}}{w_{ij'}}.$$

Thus, under hypotheses (i) and (ii) we have that the coverage rates are proportional to the weights. □

3.2.4 Optimization Model for Secondary Objectives

Since a priority group may receive more than one type of vaccine, it is possible that model (3.1) has multiple optimal solutions. Here, we consider secondary objectives that allow us to select from among these solutions. While proportionally fair coverage is the primary objective, we prefer to reduce the number of vaccine types allocated to a priority group because this helps provide clear direction to healthcare providers on which type of vaccine should be given to whom. In addition, we favor similar allocation of vaccine types across the eight health service regions, so that each region is treated similarly. We seek an optimal allocation accounting for these secondary objectives. In particular, we seek a solution that foremost achieves a proportionally fair allocation and secondarily strives for sparsity in the number of vaccine type-priority group pairs and equity in the composition of vaccine types across health service regions.

In multi-objective optimization, two common approaches to handle multiple objectives are so-called weighted-sum method and lexicographic method [16, 33]. The

weighted-sum method assigns a weight to each objective and sums them to form a single objective. By considering the units of each individual objective function and adjusting their weights, the resulting optimal solution reflects the relative importance of the objectives. On the other hand, in the lexicographic method, we optimize each individual objective one at a time in decreasing order of importance. This ensures the optimality of the most important objective and then, if there are multiple optimal solutions under the first objective, pursues the second most important objective. If a tolerance for optimality of each individual objective is allowed, the lexicographic method becomes the so-called the hierarchical method [33]. In allocating vaccines, we use the hierarchical method to ensure proportional fairness across counties. Then, we use the weighted-sum method to pick an optimal allocation which balances two secondary objectives.

We use model (3.5) to accomplish the secondary objectives while keeping the

proportional fairness within a factor of $1-\epsilon$ of g_{ij}, achieved by model (3.1):

$$\min_{Q,V,Y,\overline{Y}} \quad \sum_{(j,k)\in JK} V_{jk} + \sum_{(h,j,k)\in HJK} \frac{|Y_{hjk} - \overline{Y}_{jk}|}{|H|} \tag{3.5a}$$

$$\text{s.t.} \quad \text{(3.1b), (3.1c), (3.1d)}$$

$$f_{ij} + \frac{\sum_{k\in K_j} Q_{ijk}}{n_{ij}} \geq g_{ij}(1-\epsilon), \forall (i,j) \in IJ \tag{3.5b}$$

$$\sum_{i\in I:(i,j,k)\in IJK} Q_{ijk} \leq MV_{jk}, \forall (j,k) \in JK \tag{3.5c}$$

$$Y_{hjk} = \frac{\sum_{i\in I_h:(i,j,k)\in IJK} Q_{ijk}}{\sum_{i\in I_h:(i,j)\in IJ} n_{ij}}, \forall (h,j,k) \in HJK \tag{3.5d}$$

$$\overline{Y}_{jk} = \frac{\sum_{h\in H} Y_{hjk}}{|H|}, \forall (j,k) \in JK \tag{3.5e}$$

$$V_{jk} \in \{0,1\}, \forall (j,k) \in JK. \tag{3.5f}$$

Instead of achieving the optimal proportionally fair coverage of model (3.1), we allow the final coverage of a county-priority group to be at most ϵ away from optimality. By doing so, we provide more opportunity for model (3.5) to improve secondary objectives.

The inputs of model (3.5) include (i) the inputs of model (3.1), (ii) the optimal proportionally fair coverage rates, g_{ij}, from model (3.1), and (iii) subsets I_h, JK, and HJK. Then, the model provides an optimal allocation, denoted Q^{**}_{ijk}, and, consequently, the associated coverage of each county-priority group pair, which considers

proportional fairness and the secondary objectives.

The objective function in (3.5a) consists of two terms. The first term represents the number of vaccine type-priority group pairs that receive doses, and the second term measures the variation of vaccine types allocated across health service regions. In order to understand the second term, we fix a (j, k) pair for the moment. As defined mathematically in constraint (3.5d) and (3.5e), Y_{hjk} is the regional coverage of priority group j in region h that comes from available HSR doses of vaccine type k, and $\overline{Y}_{jk}$ is its average across all health service regions. The average absolute deviation,

$$\sum_{h \in H} \frac{|Y_{hjk} - \overline{Y}_{jk}|}{|H|},$$

represents the dispersion of vaccine type k contributing to the coverage of priority group j among regions. By summing this average absolute deviation across $j \in J$, we obtain the dispersion of vaccine type k contributing to the coverage among regions. Lastly, we sum the term across $k \in K$ to form a measure of variation of vaccine types allocated among health service regions. We could assign different weights to the two objectives in (3.5a) to adjust the relative importance of policy simplicity and regional equity. Nevertheless, we obtain desirable results, as we describe later, with equal weights.

In addition to constraints (3.1b)-(3.1d), we have five other constraints to satisfy. Constraint (3.5b) ensures that the final coverage of a county-priority group pair is no more than ϵ away from the optimal proportionally fair coverage of model (3.1). Constraints (3.5c) and (3.5f) calculate the number of vaccine type-priority group pairs that receive doses. Constraints (3.5d) and (3.5e) define Y_{hjk} and $\overline{Y}_{jk}$ for measuring the variation of vaccine types allocated among health service regions.

3.2.5 Near-Optimal Integral Allocation

By using models (3.1) and (3.5) with proper inputs, we allocate available HSR doses to eligible county-priority group pairs in a proportionally fair manner with two secondary objectives in mind. However, these two models ignore integrality of vaccine doses. Hence, we construct a post-processing step to find a near-optimal solution in which integer-valued HSR doses are allocated.

First, for each vaccine type k, we take the floor of each Q^{**}_{ijk} from model (3.5), denoted $\widehat{Q}^{**}_{ijk}$. Next, we calculate the difference between the sum of Q^{**}_{ijk} across all (i, j) pairs in subset IJ and that of $\widehat{Q}^{**}_{ijk}$. This difference is the extra doses that we have from the fractional part of Q^{**}_{ijk}. We sort eligible (i, j) pairs in descending order of their fractions and add one dose to $\widehat{Q}^{**}_{ijk}$ by this order until all the extra doses are allocated.

3.3 Data for the 2009 H1N1 Pandemic Simulation

Before using our modeling framework to simulate vaccine allocation in Texas for the 2009 H1N1 pandemic, we describe the data we use as input, including priority groups and vaccine doses distributed during the 2009 pandemic.

3.3.1 Priority Group Population Estimation

During the 2009 H1N1 pandemic, the CDC's Advisory Committee on Immunization Practices (ACIP) recommended the following groups be vaccinated with higher priority: (i) pregnant women, (ii) household contacts and caregivers for children younger than 6 months, (iii) healthcare and emergency medical services personnel, (iv) people aged 6 months through 24 years, and (v) people aged 25 through

64 years with high risk [63]. Based on availability of demographic data, we instead consider the following five priority groups: (i) people aged 0-3 years, (ii) people aged 4-24 years, (iii) people aged 25-64 years with high risk, (iv) pregnant women, and (v) infant caregivers. The details of estimating the population in each priority group are given in [25], where we estimate the population of each priority group mostly based on demographic data from the U.S. Census Bureau for 2010, using age information at one-year increments at the geographic resolution of counties. Table 3.1 lists the estimated population of these five priority groups in Texas. The total population of the five priority groups we use is about 13.5 million.

Table 3.1: Population of each priority group in Texas during the 2009 H1N1 pandemic, estimated based on U.S. Census Bureau data for 2010. See [25] for the detailed estimation procedure.

Priority group	Population
0-3 years	1,568,427
4-24 years	7,632,499
25-64 years (high risk)	3,276,939
Pregnant women	342,432
Infant caregivers	681,930
Total	13,502,227

3.3.2 Vaccine Allocation

During the 2009 H1N1 pandemic, DSHS used three channels to distribute vaccines: RPs, LHDs, and HSRs. As we mention above, the first two allocations are pull-based and the last allocation channel is push-based. In 2009, four types of vaccines were used: pre-filled syringe for baby (PFS for baby), pre-filled syringe (PFS), multi-dose vial (MDV), and live attenuated influenza vaccine (LAIV), with each having different eligible priority groups. We obtain the total number of vaccine

doses delivered to RPs in each county in the state of Texas from [39] and to each LHD and HSR from [40, 41], as shown in Table 3.2. The total number of vaccine doses distributed in Texas as of August 3, 2010 is about 8.68 million. However, the data have no resolution on vaccine types. From an internal DSHS dashboard report for H1N1 vaccines [42], we obtain the total number of each vaccine type distributed in Texas on a weekly basis, but it has no resolution on where these vaccines were allocated. Table 3.3 lists the percentage of total doses distributed as of January 29, 2010 for each vaccine type from the report.

Table 3.2: Vaccine doses allocated to RPs, LHDs, and HSRs in the 2009 H1N1 pandemic as of August 3, 2010 [39–41].

	Doses	Percentage (%)
RPs	6,676,310	77
LHDs	1,419,540	16
HSRs	590,380	7

Table 3.3: Percentage of total doses distributed as of January 29, 2010 for each vaccine type used in the 2009 H1N1 pandemic [42].

Vaccine type	Percentage (%)
PFS baby	3
PFS	17
MDV	60
LAIV	20

In order to estimate the doses of each vaccine type delivered to a county via RPs, we assume these doses have the same composition as that of the total doses distributed in the whole state, i.e., PFS for baby 3%, PFS 17%, MDV 60%, and LAIV 20%. We further assume the doses allocated to each LHD have the same composition. If an LHD serves more than one county, the LHD is assumed to distribute its doses to its counties in proportion to county populations. Finally, for HSR doses we assume

they have the same composition as that of the total doses distributed in the whole state as well, and DSHS has the control to allocate any integral number of doses to a county and, in turn, to a priority group.

The suitability of vaccine types for each of the five priority groups is listed in Table 3.4. In particular, the PFS baby vaccine can only be used for group (i). The PFS and MDV vaccine types can be given to those in all priority groups, except group (i). The LAIV type is suitable for groups (ii) and (v). In order to obtain the coverage of each county-priority group pair prior to the allocation of HSR doses, we assume that within a county, RPs and LHDs allocate vaccines of a certain type to these five priority groups in proportion to the population of the priority group, if the group is suitable for the vaccine type. As a result, if one priority group is eligible for two types of vaccines, the group would have a higher coverage rate (via RP and/or LHD doses) than another group eligible for only one of the two types.

Table 3.4: Suitability of vaccine types for each priority group: 1 indicates that a vaccine type is suitable for a priority group and 0 indicates it is not.

	PFS baby	PFS	MDV	LAIV
0-3 years	1	0	0	0
4-24 years	0	1	1	1
25-64 years (high risk)	0	1	1	0
Pregnant women	0	1	1	0
Infant caregivers	0	1	1	1

We define an *ideal ratio* as the ratio of available doses of all types allocated to an area to the area's population of all priority groups. Assuming all county-priority group pairs have equal priority, and ignoring vaccine suitability, the ideal ratio is the most equitable coverage we could achieve for all pairs simultaneously. The vaccine doses distributed in the 2009 H1N1 pandemic have the ideal ratio of 64.3%

(= 8.68/13.50). The actual coverage rate of each county-priority group pair might vary widely due to priority group population, vaccine suitability, and uneven RP and LHD allocations. Nevertheless, we can use 64.3% as a reference for proportionally fair coverage.

3.4 Results

In 2009, DSHS reserved 7% of total SNS doses for allocation to HSRs for the purpose of boosting the coverage of counties where an insufficient number of doses were distributed to RPs. As we mention above, there are 189 counties qualified for HSR doses. We calculate the ideal ratio for two groups of counties, the rural counties served by HSRs and the urban counties served by LHDs. The rural counties have an ideal ratio of 61.9% when we include the 7% of HSR doses along with their RP doses. The urban counties have an ideal ratio of 64.8% based on their RP and LHD doses. So, while the rural counties fall a bit shy of the state-wide ratio of 64.3%, we see that the 7% of HSR doses helps achieve a more equitable ideal ratio for the rural counties, i.e., up from 30.2% without HSR doses to 61.9% with HSR doses. This suggests the possibility of having similar coverage across the 254 counties in Texas by properly allocating available HSR doses. Further analysis of the portion of total doses reserved by DSHS for discretionary allocation to HSRs is discussed in Section 3.4.2 in the context of sensitivity analysis.

We use our optimization framework guided by models (3.1) and (3.5) to estimate the performance of the 7% of discretionary doses that DSHS can allocate to HSRs, using data available from the 2009 H1N1 pandemic as described in Section 3.3. Primarily, we focus on achieving proportional fairness, although we also consider the

two secondary objectives that we describe in Section 3.2.4.

3.4.1 2009 H1N1 Pandemic Simulation

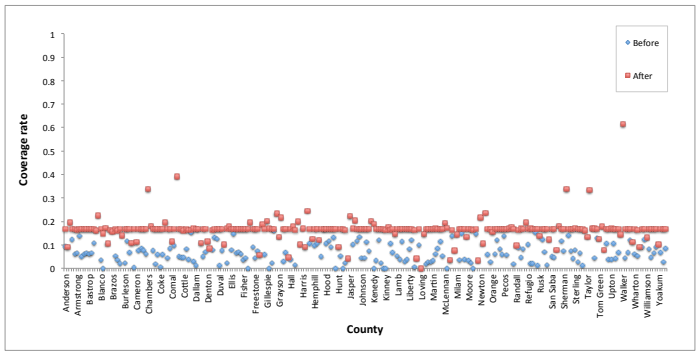
In 2009, doses were delivered to Texas from the SNS on a weekly basis and, in turn, allocated across the state. However, for simplicity of exposition, we speak of a one-time allocation in our 2009 H1N1 pandemic simulation; i.e., we assume all the RP and LHD doses are allocated to county-priority groups and all 7% of total doses reserved for HSRs are available for allocation. Furthermore, we set the weight for each county-priority group pair to $w_{ij} = 1$ because the ACIP did not distinguish among the five priority groups. The purpose of this analysis is to see to what extent the HSR doses (7% of all doses) can achieve proportional fairness in vaccine allocation.

Figure 3.2 shows the coverage for each priority group and all groups aggregated for each of the 254 counties in Texas. The blue dots represent the coverage before allocation of the HSR doses, and the red dots represent the coverage after allocation of the HSR doses. We can see that, in large part, the HSR doses bring the under-served priority groups in the counties qualified for the HSR doses, up to the same level. In particular, the HSR doses bring the coverage of the priority group of 0-3 years to at least 17%, and the coverage for all other priority groups to 64%. The reason that the priority group of 0-3 years differs is that this population is only eligible for the PFS baby vaccine type and we have a relatively small number of PFS baby doses available (3% of total doses as shown in Table 3.3). There are some red dots under the proportionally fair coverage because these counties are among the 65 served by LHDs and not qualified for HSR doses. There are some red dots above the proportionally fair coverage because these counties received an excess number of doses, relatively to the proportionally fair rate, via RPs and/or LHDs. Figure 3.2(f) shows similar

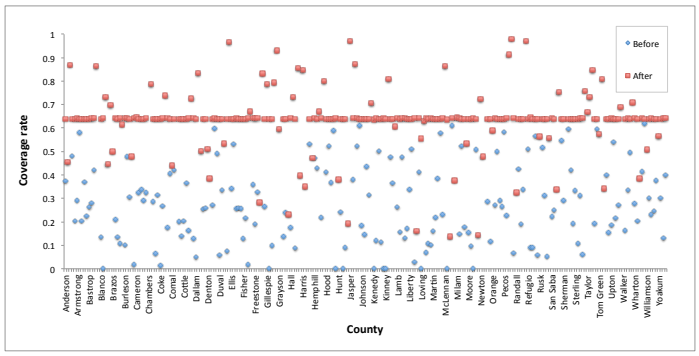
before-and-after results for the county coverage rate by aggregating all of the priority groups.

As we mention above, there may be more than one optimal allocation of HSR doses that achieves proportionally fair coverage because a priority group can receive multiple types of vaccines. Rather than arbitrarily choosing among optimal allocations, we select one that has fewer priority group-vaccine type pairs that receive doses and that maintains equity of vaccine types across health service regions. Table 3.5 summarizes two such allocations that achieve the same level of proportional fairness (to three significant digits, i.e., $\epsilon = 0.001$ in model (3.5)). However, the first solution only considers proportional fairness while the second solution also considers the sparsity objective for having fewer priority group-vaccine type pairs and for maintaining equity of vaccine types across the health service regions. The table illustrates the effect of the former secondary objective.

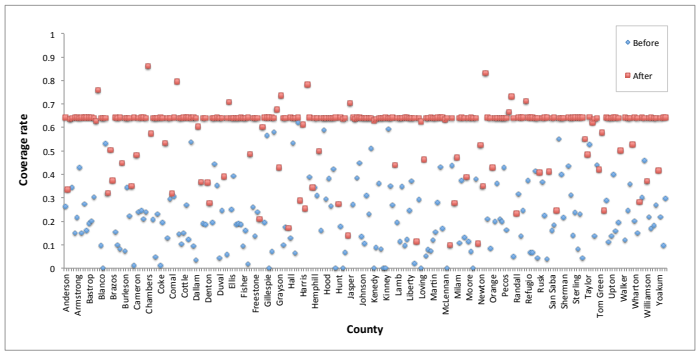
From Table 3.5, we can see that there is no difference for the priority group of 0-3 years because there is a one-to-one matching between this priority group and the PFS baby vaccine type. However, for the other four priority groups, the allocations differ. In the solution of part (a) in the table, three types of vaccines are allocated to the largest priority group (4-24 years) while two types of vaccines are used in the solution of part (b). The smaller priority groups of pregnant women and infant caregivers drop from two and three types of vaccines to one type of vaccine. The group of high-risk people aged 25-64 years receives PFS and MDV vaccines in both solutions. Relative to the solution of part (a), the solution shown in part (b) may simplify policy recommendations issued to healthcare providers on the type of vaccines to provide to each priority group. We emphasize that the changes in these two solutions are



(a) 0-3 years

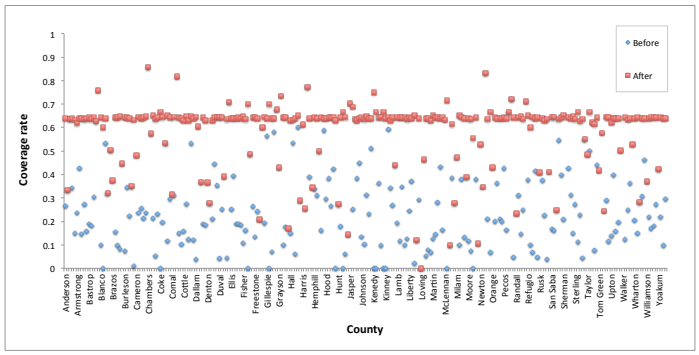


(b) 4-24 years

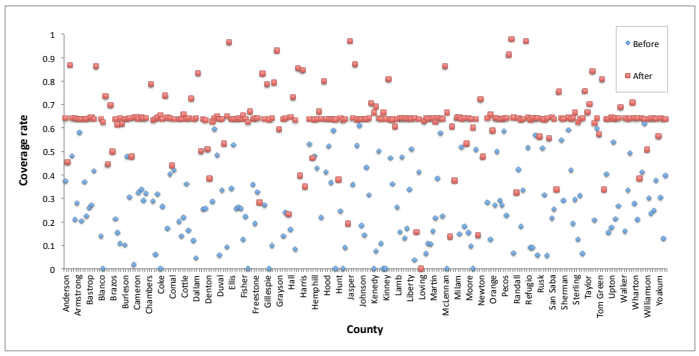


(c) 25-64 years (high risk)

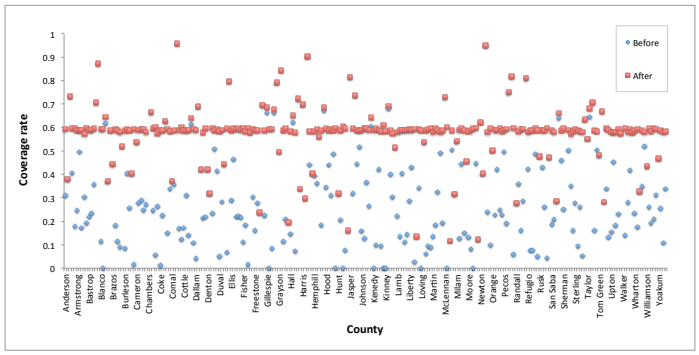
Figure 3.2: Coverage at county level for each priority group before (blue) and after (red) HSR doses are allocated. The sub-captions indicate the priority groups. The x-axis has all 254 counties in Texas, in alphabetical order, even though only a subset of the counties are listed, and even though HSR doses are allocated to only 189 out of the 254 counties.



(d) Pregnant women



(e) Infant caregivers



(f) All priority groups aggregated

Figure 3.2 (cont.): Coverage at county level for each priority group before (blue) and after (red) HSR doses are allocated. The sub-captions indicate the priority groups. The x-axis has all 254 counties in Texas, in alphabetical order, even though only a subset of the counties are listed, and even though HSR doses are allocated to only 189 out of the 254 counties.

obtained with no loss in the proportional fairness criteria (to three significant digits).

Table 3.5: HSR doses allocated to the priority groups by vaccine type. Solutions are expressed as a percentage of doses assigned to each priority group. In the solution of part (a), we only consider proportional fairness, and in the solution of part (b) we also simultaneously account for two secondary objectives: sparsity of vaccine type-priority group pairs and equity of vaccine allocations across health service regions. The differences in the two solutions illustrate the sparsity issue.

(a) Without considering secondary objectives

Percentage (%)	Doses	PFS baby	PFS	MDV	LAIV
0-3 years	17,708	100	0	0	0
4-24 years	343,236	0	12.6	55.7	31.7
25-64 years (high risk)	187,155	0	22.8	77.2	0
Pregnant women	15,742	0	44.2	55.8	0
Infant caregivers	26,539	0	28.2	37.2	34.6
Total	590,380				

(b) Considering secondary objectives

Percentage (%)	Doses	PFS baby	PFS	MDV	LAIV
0-3 years	17,708	100	0	0	0
4-24 years	343,238	0	0	73.3	26.7
25-64 years (high risk)	187,162	0	45.2	54.8	0
Pregnant women	15,736	0	100	0	0
Infant caregivers	26,536	0	0	0	100
Total	590,380				

Table 3.6 compares the same two solutions shown in Table 3.5, except that we now display the percentage of doses allocated to each region by vaccine type. Again, the allocations to the priority group of 0-3 years are identical for the reason we discuss above. Overall, the variability of the solutions across the health service regions is decreased in the solution of part (b) relative to that of part (a). In particular, the allocations of the PFS and MDV vaccines are less variable in part (b)'s solution, although the variability of the LAIV allocation has increased somewhat.

Table 3.6: HSR doses allocated to regions by vaccine type. Solutions are expressed as a percentage of doses assigned to each region. In the solution of part (a), we only consider proportional fairness, and in the solution of part (b) we also simultaneously account for two secondary objectives: sparsity of vaccine type-priority group pairs and equity of vaccine allocations across health service regions. The differences in the two solutions illustrate the issue of equity among health service regions. See Figure 2.1 for a map of Texas with the regions we label in the first column.

(a)Without considering secondary objectives

Percentage (%)	PFS baby	PFS	MDV	LAIV
HSR 1	3.5	26.3	46.2	24.0
HSR 2/3	2.8	12.8	66.3	18.1
HSR 4/5N	2.8	18.9	57.2	21.1
HSR 6/5S	3.0	13.8	62.2	21.0
HSR 7	2.9	16.6	60.2	20.3
HSR 8	3.1	16.8	61.1	19.0
HSR 9/10	3.3	31.1	44.2	21.4
HSR 11	4.3	21.3	52.9	21.5

(b) Considering secondary objectives

Percentage (%)	PFS baby	PFS	MDV	LAIV
HSR 1	3.5	14.5	57.2	24.8
HSR 2/3	2.8	18.1	60.4	18.7
HSR 4/5N	2.8	17.5	62.7	17.0
HSR 6/5S	3.0	14.1	64.3	18.6
HSR 7	2.9	16.1	55.1	25.9
HSR 8	3.1	19.0	63.2	14.7
HSR 9/10	3.3	19.2	54.2	23.3
HSR 11	4.3	13.2	59.9	22.6

3.4.2 Sensitivity Analysis

In addition to simulating the vaccine allocation during the 2009 H1N1 pandemic in Texas using historical distribution data, we perform a sensitivity analysis on the 7% of total doses available for HSR allocation at DSHS's discretion, ranging from 1% to 13% in increments of 2%. As we increase the portion of doses that DSHS reserves for discretionary allocation to HSRs, we proportionally decrease doses allocated based on RP and LHD requests. Increasing this percentage leads to better allocations to the 189 rural counties covered by HSRs, but leads to worse allocations to the 65 urban counties covered by LHDs. Table 3.7 shows the ideal ratios of the 189 rural counties served by HSRs and the 65 urban counties served by LHDs under different portions of total doses reserved for HSRs. The ratio is the ideal coverage that we could achieve for all county-priority group pairs simultaneously, ignoring vaccine suitability.

Table 3.7: Ideal ratios (%) of the rural areas and the urban areas under different portions of total vaccine doses reserved for HSR allocation. The rural areas include the 189 counties served by HSRs, and the urban areas include the other 65 counties served by LHDs. The base case of 7% is indicated in bold font.

Percentage of total vaccines reserved for HSRs (%)	189 rural counties served by HSRs (%)	65 urban counties served by LHDs (%)
1	38.1	68.9
3	46.0	67.5
5	54.0	66.1
7	**61.9**	**64.8**
9	69.9	63.4
11	77.8	62.0
13	85.8	60.6

As we mention in Section 3.3.2, reserving 7% of the doses for discretionary allocation to HSRs results in an ideal ratio of 61.9% in the rural counties and 64.8%

in the urban counties. Table 3.7 shows the ideal ratio of the rural counties changes more quickly than that of the urban counties as we vary the reservation percentage. This is because the total population of the priority groups in the rural counties is smaller than in the urban counties (2.01 million vs. 11.49 million). If the portion of total vaccines reserved for allocation to HSRs drops to 5%, the ideal ratios differ more than 10%. On the other hand, if the portion goes up to 9%, the rural areas (189 counties) may have better coverage than the urban areas (65 counties) by more than 6%. Hence, 7% seems to be a good portion in terms of having equal coverage across the 254 counties, based on the ideal coverage calculation.

We now turn to results obtained using our optimization framework. Table 3.8 shows the median coverage (before-and-after HSR allocation) for all priority groups aggregated for the 189 rural counties. We use median instead of mean to represent the central tendency of the coverage rates for following reasons. First, the distribution of coverage rates for the 189 rural counties after the allocation of HSR doses is highly skewed to right because all of them have at least the proportional fair coverage and some did well under RP requests. The mean, which is not weighted by population size, is highly influenced by the counties of small size. On the other hand, a population-weighted mean is very close to the ideal ratio in Table 3.7. Second, as the result of the optimal allocation, the under-served counties have a similar final coverage, which can be represented by the median.

Table 3.8 indicates that the median before HSR allocation decreases about 0.4-0.6% for every 2% increment of the reserved portion. On the other hand, we see from the table that the median after HSR allocation increases in a nonlinear manner as the reserved portion grows. The value of 60.2% for the 7% row in Table 3.8 is smaller

than the 61.9% in Table 3.7 because the latter include counties whose allocations exceed the proportionally fair level due to RP allocations.

We emphasize that even though our optimization framework aspires to achieve a proportionally fair allocation for most county-priority group pairs, this may not be achievable. Previously allocated doses, from RPs and LHDs, may be so imbalanced that the vaccine doses available for HSR allocation cannot provide proportional fairness for most counties. For example, if 1% of total doses are discretionary, we bring 64 out of the 189 rural counties from their prior coverage rates up to 27.5%. If the portion is 7%, then our HSR allocation can achieve proportionally fair coverage in 118 out of the 189 counties. Furthermore, we can see from Table 3.8 that 7% is around the point that the increments of the median start to stabilize, which suggests that 7% achieves most of the benefit of equitability we can obtain from HSR allocation.

In addition, we use a boxplot to visualize the variation in coverage among the 189 rural counties. Figure 3.3 shows boxplots with whiskers from the minimum to the maximum of aggregated coverage for the 189 rural counties before-and-after HSR allocation. Comparing part (a) and part (b) in the figure, we can see that the HSR doses effectively shrink the distances between the minimum, first quartile, median, and third quartile. The maximum coverage rates among the 189 rural counties are the same in part (a) and part (b) since we do not allocate HSR doses to over-served counties. Also, the maximum becomes smaller with the growth of the portion since we take doses from RPs in proportion to their requests. On the other hand, we can see that the HSR doses bring the minimum coverage up to nearly the same level as the median. The gaps between the minimum and median are because the discretionary portion is small or because we cannot allocate fractional doses. Moreover, we

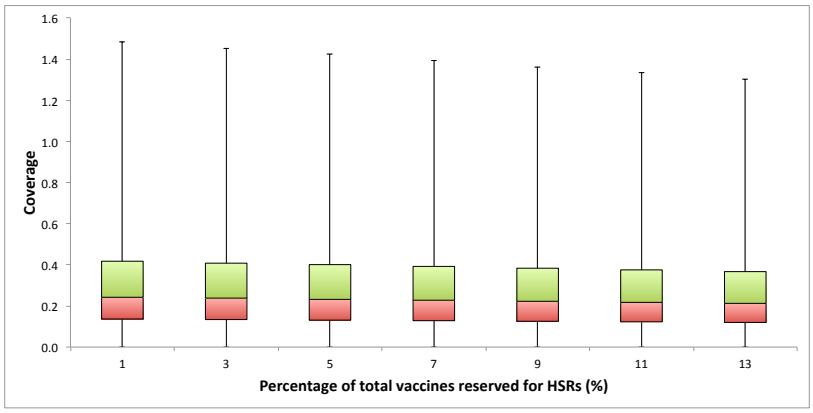
can see that the variation of aggregated coverage decreases with the growth of the discretionary portion since the box size and the distance between whiskers becomes smaller from the left to the right in Figure 3.3(b). In particular, we see that when the portion is 3% or larger, the first quartile, median and third quartile have the same value, indicating most of the counties have the same coverage.

Table 3.8: Median (%) of the coverage of all priority groups aggregated for the 189 rural counties before-and-after HSR allocation under different portions of total vaccines reserved for HSRs. The base case is 7%, indicated in bold font.

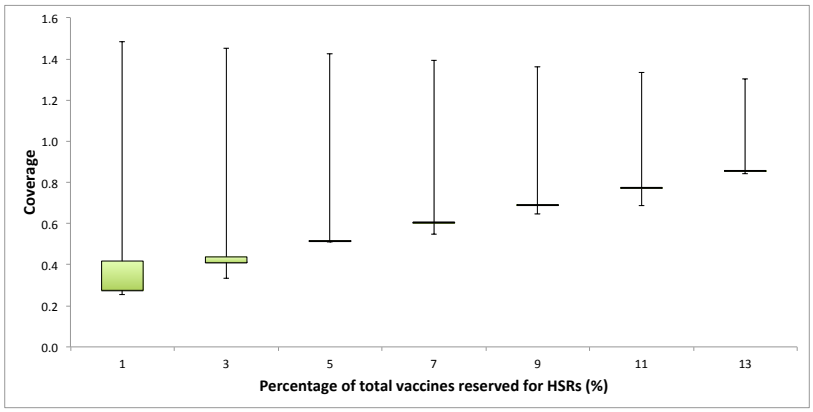
Percentage of total vaccines reserved for HSRs (%)	Before HSR allocation	After HSR allocation
1	24.2	27.5
3	23.8	40.9
5	23.2	51.2
7	**22.8**	**60.2**
9	22.3	68.7
11	21.7	77.1
13	21.3	85.3

3.5 Discussion

Our optimization framework aims to bring under-served priority groups up to a proportionally fair coverage, using available HSR doses. However, our approach may not effectively shrink the gap between over-served and under-served priority groups if there is a large imbalance from previously allocated doses as we see in the sensitivity analysis of Section 3.4. Furthermore, based on the number of available doses, and the mapping between vaccine type-priority group pairs, it may be impossible to achieve proportional fairness between different priority groups. Some priority groups may simply have fewer doses for which they are eligible than other priority groups as we see in the results for the 0-3-year-old priority group.



(a) Before HSR allocation



(b) After HSR allocation

Figure 3.3: Boxplot with whiskers from the minimum to the maximum of aggregated coverage for the 189 rural counties under different portions of total vaccines reserved for HSRs. The two boxes represent the first quartile to the median (red) and the median to the third quartile (green) while the whiskers show the minimum and maximum.

Our optimization-based approach to proportional fairness is also capable of addressing the relative importance of county-priority group pairs. By assigning different weights to county-priority group pairs (where the least important pair is assigned unit weight), the model seeks an allocation achieving equal coverage when weighted by their relative importance. For example, if one county-priority group pair has twice the weight of another, we seek twice the coverage rate for the higher priority pair.

For demonstrating the capability of our optimization framework, we simulate the 2009 H1N1 vaccine allocation in Texas under the assumption of a one-time distribution. The framework can also be applied in a time-dynamic rolling-horizon manner. In such a setting, previously allocated vaccines doses will include RP and LHD doses and will also include HSR doses allocated in previous time periods. The optimization models then seek equal coverage of each county-priority group pair at each time period.

In our analysis here, we have not attempted to account for the potential benefit of geographic allocation of vaccines according to the time-dynamic spread of influenza. However, our framework could be used to allocate available doses in this manner by using the optimization models on, say, a weekly basis and assigning different weights to county-priority group pairs to account for the spread of the disease.

Finally, we consider neither the potentially different costs of distributing vaccines via RPs, LHDs, and HSRs, nor the uptake of doses assigned to these three channels. That is, we assume one dose would reach the public with the same uptake at the same cost, regardless of distribution channel. Doses distributed via RPs might be more accessible since people are more aware of their local healthcare providers, compared to the temporary medical resource points of distribution that could be set

up by LHDs and HSRs. The tradeoff among distribution cost, dose uptake, and equitable vaccine coverage needs further study.

Chapter 4

Effect of Demand Aggregation and Bin Delivery on an In-Plant Just-in-Time Parts Supply System

4.1 Introduction

Fasteners, such as washers, bolts, and nuts are used to affix objects together, e.g., thousands of components of an engine. The ABC analysis used to categorize inventoried parts [21] categorizes these parts, usually small, inexpensive and of high demand, as class C parts and often suggests purchasing them in bulk. Due to extra labor involved in repacking, an engine assembly plant may want to deliver these parts from its warehouse to assembly workstations in bins (or boxes), as they were shipped from the suppliers, instead of opening a box and delivering the exact amount needed to the workstations. However, bin delivery may result in not having the part at the right workstations when needed since all of the inventory is allocated to other workstations, which may cause the whole assembly line to shut down.

Moreover, risk pooling has been implemented in several areas to reduce necessary resources for achieving a predefined performance level, especially in supply-chain management. The idea of risk pooling is to combine several stochastic demands to reduce the total variation with the cost of dispensing resources from the central warehouse to local points of use after the local demands are realized. This chapter is motivated by a project within an engine assembly plant located in the state of Texas. The plant uses a manufacturing execution system (MES) to manage and replenish

such class C parts, where it calculates the order quantity based on the aggregation of next-day demand across all workstations. After receiving the parts, the plant stores them in the warehouse first and delivers them to workstations in bins when the line-side inventory at a workstation is below a predefined threshold. In this chapter, we implement the risk pooling idea to derive a short-cut formula for the extra inventory needed to control the risk of not satisfying all workstation demands due to demand aggregation and bin delivery.

Here, we review the literature on in-plant material supply, lot sizing and risk pooling. In general, there are two types of in-plant material supply policies, namely kitting and continuous supply [22, 31]. Continuous supply is also called line-side stocking or kanban-based just-in-time (JIT) continuous supply [10, 19]. In kitting, we pre-pick a set of components into a kit container according to the assembly operations performed to an end product at one or several workstations. On the other hand, continuous supply usually uses a two-bin storage and replenishment system for each component at workstations [22, 31]. The system delivers another bin of the component to a workstation when the workstation empties a bin. Moreover, there are several transportation and material handling methods to move components inside a plant, including using pallet jacks, push carts, tuggers and trains [7]. For continuous supply, a milk run is a common transportation method where a tugger driver drives a train from the warehouse, carrying several requested components to visit multiple workstations periodically [31]. In this chapter, the plant studied uses a continuous supply policy and milk runs to replenish class C parts.

Several researchers have studied the advantages and disadvantages between kitting and continuous supply. Hua and Johnson [22] identify a number of research

issues that might influence the choice between kitting and continuous supply at an electronics assembly plant, including product characteristics, storage, material handling, etc. They state that these research issues are worth further investigation and once these issues are addressed, a methodology or tool could be developed to assist a company in deciding which system to use. Hanson and Brolin [19] also conduct a comparison of kitting and continuous supply in in-plant materials with a study of two cases within the Swedish automotive assembly industry. Performance indices include man-hour consumption, product quality, flexibility, inventory levels, and space requirements. They list relative effects of these two material supply policies based on interviews with the corresponding personnel. Limère et al. [31] introduce a mathematical cost model to compare kitting and continuous supply at an automotive company. The cost model considers the average yearly labor cost for picking at the assembly line, internal transportation cost, kit assembly operation, and replenishment cost. They conclude that neither kitting nor continuous supply dominates in all parts. Instead, hybrid policies where some parts use kitting and others use continuous supply perform better. In addition to distinguishing between kitting and continuous supply, Caputo and Pelagagge [10] further classify continuous supply as periodic inventory review or continuous inventory monitoring. The difference between these two policies is the frequency a line-side stock is replenished. Then, a descriptive model, considering work in process, holding cost, equipment, workforce requirements, as well as intensity of containers flows, is developed for evaluating these three policies. They obtain a similar conclusion that a hybrid feeding policy performs better than a single feeding policy common to all components. As the above researchers suggest, the plant studied in this chapter uses a continuous supply policy to manage and replenish class C parts since they are usually small, inexpensive, and of high demand. However,

none of them discuss how much inventory should the plant hold explicitly.

In terms of lot sizing study in continuous supply, Hanson and Finnsgård [20] investigate the impact of the unit load size on in-plant material supply efficiency. Smaller unit loads can reduce the time the assembler takes to reach the components while larger unit loads require fewer moves for a given volume of components. They find that the increased delivery frequency required for smaller unit loads does not necessarily increase man-hours due to the savings on the improved component presentation at workstations which requires fewer assemblers. Battini et al. [6] study a mixed-models assembly system and use a hierarchical structure to make two kinds of decisions: (i) deciding on centralized or decentralized stocking for a part, and (ii) if a centralized policy is preferred, deciding which feeding policy from the warehouse to workstations the plant should implement: pallet, trolley, or kit. Pallet and trolley are two continuous supply methods, the difference between them is the volume and timing of parts delivery to workstations. Neither of the two studies above discusses the amount of inventory the plant should hold.

Finally, risk pooling has been applied in inventory management to reduce the necessary inventory for a given risk level. The idea is that by aggregating stochastic demands, we can reduce the total variation drastically, especially when the demands are highly negatively correlated. We can apply risk pooling to aggregate across customers of a certain product and/or across products with common components. See [49] for a detailed risk pooling discussion and concrete examples. In particular, the plant we study produces customized engines that have a large number of common components, especially class C parts. As a consequence, the plant may use a class C part at many workstations. We do not apply the risk pooling idea to workstation

demands directly when the workstation demands are known before ordering. Instead, we consider the remaining inventory at a workstation (due to bin delivery) as a random variable and utilize the risk pooling idea to estimate extra inventory needed to cover the effect of bin delivery.

The organization of the remainder of this chapter is as follows. In Section 4.2, we describe the simplified existing replenishment process used in the plant, discuss the effect of demand aggregation and bin quantities, and categorize four cases to study according to information availability in the order quantity calculation. In Section 4.3, we build a stylized model for the simplified replenishment process, discuss the special case of a single workstation with sufficient supply, and for each case derive a short-cut formula estimating the extra inventory needed for a given risk level. In Section 4.4, we use numerical examples to validate the formulae and perform sensitivity analyses on workstation demand variation and bin size. We then further discuss the performance of the short-cut formulae and the tradeoff between extra inventory needed and risk level in Section 4.5.

4.2 Replenishment Process

The replenishment process we discuss in this chapter is simplified from the existing process in the engine assembly plant. A class C part may be used in several workstations. The daily demand of each workstation can be different and is known one day in advance. Every day in the afternoon, for each part the MES calculates the order quantity in pieces for the next day based on the aggregated next-day demand with the goal of keeping at least a certain amount of inventory (called minimum inventory) at the end of each day. To be more specific, the MES calculates the order

quantity by aggregating next-day demands across all workstations plus the minimum inventory minus the (predicted) existing inventory at the end of the day, including inventory in the warehouse and that at workstations (line-side inventory). By doing so, the plant ensures that there is at least the minimum inventory of the part in pieces in the plant.

After the order quantity calculation, the plant rounds up to the closest bin quantity, and then sends the final order quantity to the supplier since the supplier can only ship a multiple of bins to the plant. Upon receiving the supply the following morning, the plant restores the supply in its warehouse and then delivers a multiple of bins to a workstation whenever the inventory status of the workstation is projected to go below zero in an hour, to try to maintain nonnegative inventory at the workstation. If the bin size is one piece and the minimum inventory is set to zero, every request from workstations will be satisfied since we order exactly what is needed for each day and deliver the exact amount of the part needed to a workstation. However, in reality it is not uncommon to have an excessive amount of inventory at one workstation while another workstation suffers from part unavailability because the bin size is more than one piece and all of the plant warehouse inventory has been allocated to other workstations, which may cause an eventual shut-down of the whole assembly line.

The parameter that the plant has control over is the minimum inventory and the performance metric the plant uses is the part availability in the warehouse (called warehouse part availability in the sequel), i.e., if a workstation requests a bin to deliver, there is a bin in the warehouse for the request. Given demand aggregation during the order quantity calculation, the main function of the minimum inventory is

to cover the effect of bin delivery from the warehouse to workstations. If we increase the minimum inventory, the warehouse part availability will increase. Of course, this conflicts with lowering overall inventory level.

Effect of Demand Aggregation and Bin Delivery

As we describe before, for a class C part on a given day, a workstation may suffer from not having the part when the workstation needs it while the plant overall has sufficient inventory to cover all part demand (at other workstations). The main reason is that the MES orders the part based on its aggregated next-day demand across all workstations and delivers the part to workstations in bins, not in pieces. Figure 4.1 shows a simple example illustrating the effect.

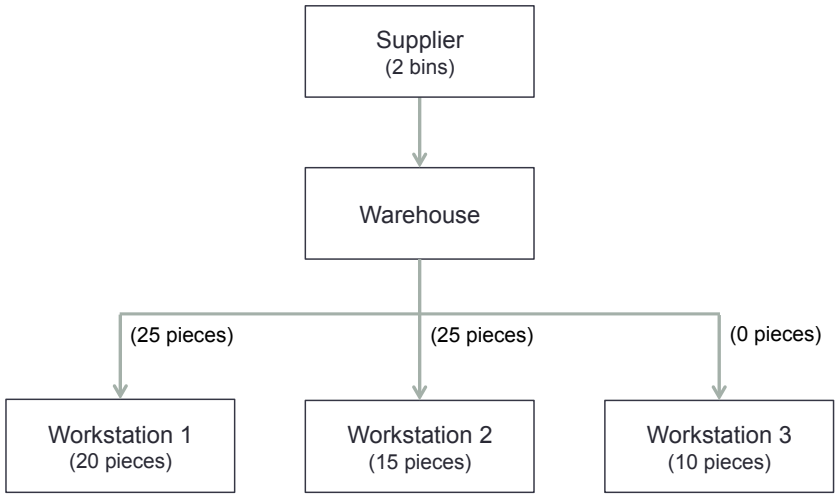


Figure 4.1: Effect of demand aggregation and bin delivery. In this small example, the bin size is 25 pieces and there are three workstations with total (aggregated) demand of two bins. Given delivering the part to workstations in bins, the warehouse can then only satisfy two out of three requests from these workstations.

Assume that the bin size is 50 pieces and there are three workstations requiring

20, 15 and 10 pieces of the part for the next day, respectively. The aggregated demand across all of these workstations is 45 pieces so that the MES will order two bins. Thus, we can only fulfill two out of three requests from the workstations since there are only two bins in the plant warehouse and we can only deliver them to workstations in bins.

Cases of Information Availability in the Order Quantity Calculation

In addition to assuming that we know next-day demand and line-side inventory in the order quantity calculation, we extend the study of the simplified replenishment process to other cases of information availability. To be more specific, we consider four cases with two-dimensional uncertainty: next-day demand and line-side inventory. Table 4.1 lists the four cases and the information we use in the order quantity calculation.

Table 4.1: Four cases of information availability in the order quantity calculation. For example, in Case 1 we know next-day demand and remaining line-side inventory and use them when calculating the order quantity.

Case clarification	Known line-side inventory	Unknown line-side inventory
Known next-day demand	Case 1	Case 2
Unknown next-day demand	Case 3	Case 4

4.3 Modeling Framework

In this section, we build a stylized model to analyze the simplified replenishment process in order to get insights into the relationships among the minimum inventory, warehouse inventory, line-side inventory, and warehouse part availability (see below for mathematical definitions). With some theory and approximations, we

further derive a short-cut formula for each case, estimating the minimum inventory needed to achieve the fill rate of workstation requests.

4.3.1 Stylized Replenishment Process Model

4.3.1.1 Model Assumptions

The model we build and the analysis we carry out in this chapter have the following assumptions:

1. The plant can only order a class C part in bins from the supplier and deliver it in bins from the warehouse to requesting workstations.

2. The plant sends the order request to the supplier at the end of a day and the supply arrives the next morning before workstations start working. The plant calculates the order quantity according to the minimum inventory, existing warehouse inventory, and/or next-day demand of workstations, and/or existing line-side inventory, depending on the setting.

3. Each workstation sends at most one request to the warehouse based on its daily demand and its line-side inventory status. The workstation computes the requested quantity so that the resulting line-side inventory is nonnegative and no more than one whole bin.

4. We assume that the order of workstations sending requests to the warehouse is random every day and the warehouse delivers the part to workstations in bins according to this order and the existing warehouse inventory.

5. The warehouse cannot send a bin to a workstation once its inventory status becomes zero.

6. We allow backorders at workstations, i.e., if the warehouse cannot fulfill a request, the line-side inventory of the workstation becomes negative and the request is backordered.

7. Once the warehouse delivers a bin to a workstation, the warehouse dedicates it to the workstation, i.e., there is no part-sharing among workstations.

In reality, the line-side inventory at a workstation is continuously monitored and the workstation sends out requests to the warehouse in a just-in-time manner, i.e., whenever the line-side inventory status is projected to go below zero in an hour. So, it is possible for a workstation to send more than one request to the warehouse in a day. In addition, if the line-side inventory status of a workstation becomes negative and the warehouse has no inventory, the plant will move inventory from other workstations to the starving one manually. Our simplifications allow us to get insights into the real replenishment process without loss of tractability.

4.3.1.2 Model Notation

We use the following notation in this chapter.

Indices and Sets

$i \in I$: workstations, $I = \{1, 2, \ldots, n\}$

$t \in T$: days, $T = \{1, 2, \ldots, m\}$

Variables and Parameters

D_i^t	:	demand (in bins) of workstation i on day t
O^t	:	order quantity (in bins) that arrives in the warehouse at the beginning of day t
Z^t	:	warehouse inventory status (in bins) at the end of day t
S_i^t	:	line-side inventory status (in bins) at workstation i at the end of day t
d_i^t	:	demand (in pieces) of workstation i on day t
s_i^t	:	line-side inventory status (in pieces) at workstation i at the end of day t
b	:	bin size
μ_i	:	mean of normally distributed daily demand (in bins) of workstation i
σ_i^2	:	variance of normally distributed daily demand (in bins) of workstation i
α	:	target warehouse part availability level, e.g., 0.95
$\bar{\alpha}$	:	average of simulated warehouse part availability estimates
γ	:	maximum relative gap (see below for mathematical definition)
$\Phi^{-1}(\cdot)$	:	inverse cumulative distribution function of the standard normal distribution

Decision Variable

Y	:	decision variable (in units of bins), whose meaning depends on the setting of O^t

For notational simplicity, we also denote $D^t = \sum_{i \in I} D_i^t$, $S^t = \sum_{i \in I} S_i^t$, $\mu = \sum_{i \in I} \mu_i$, and $\sigma^2 = \sum_{i \in I} \sigma_i^2$.

4.3.1.3 Model

Figure 4.2 displays the order of variables over time. At the end of day $t-1$, we observe the inventories in the warehouse (Z^{t-1}) and at workstations (S^{t-1}). We calculate the order quantity (O^t) according to the minimum inventory (Y), existing warehouse inventory (Z^{t-1}), and/or existing line-side inventory (S^{t-1}), and/or

aggregated next-day workstation demand (D^t), depending on the setting.

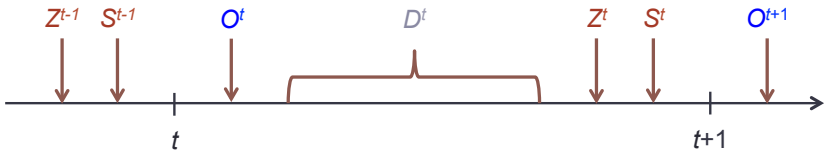


Figure 4.2: System dynamics of variables observed over time. First, we observe the existing warehouse inventory (Z^{t-1}) and line-side inventory (S^{t-1}) at the end of day $t-1$. Then, we calculate the order quantity (O^t) according to information availability. Demand on day t (D^t) happens after we receive the order quantity (O^t) and before we check the remaining warehouse inventory (Z^t) and line-side inventory (S^t).

Figure 4.3 illustrates the stylized replenishment process model for one part. At the end of day $t-1$, the plant sends out the request (O^t) to the supplier. The next morning, the supply arrives before workstations start working and the plant stores it in the warehouse first. Upon receiving their requests, the warehouse delivers the part in bins to workstations. At the end of day t, we obtain the line-side inventory status at workstation i (S_i^t) by computing the line-side inventory of the previous day (S_i^{t-1}) minus the workstation demand (D_i^t) plus the delivery amount from the warehouse.

Warehouse Part Availability

We define warehouse part availability as the fraction of days that all workstation requests are satisfied, i.e., all of the line-side inventory statuses at the end of the day are nonnegative. If the warehouse part availability is 99%, it means that within 100 days, there is (only) 1 day that at least one of the workstation requests cannot be satisfied. This performance index does not capture the number of unsatisfied workstations or the shortage of an unfulfilled request.

Order Quantity Calculation

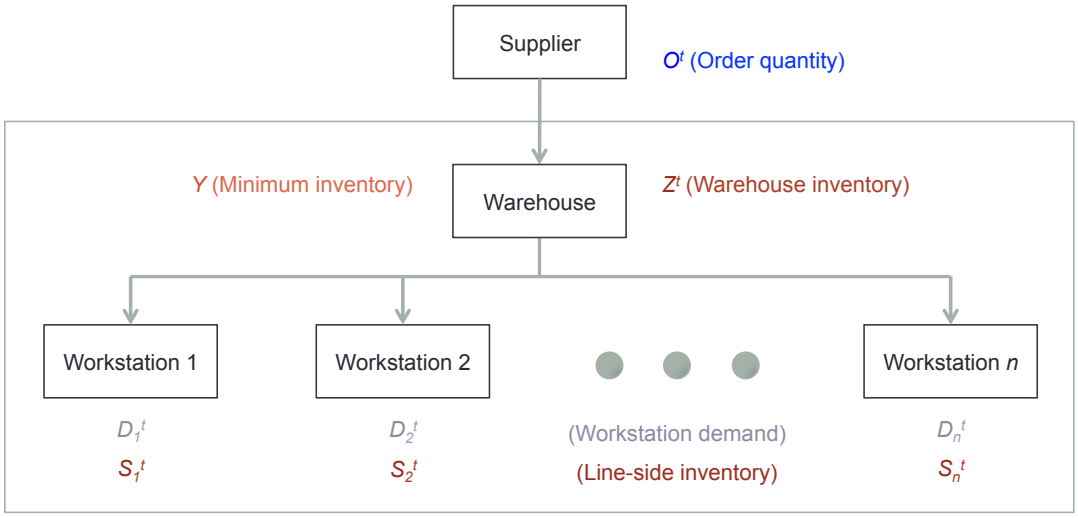


Figure 4.3: Stylized replenishment process model. The plant receives the order quantity (O^t) from the supplier at the beginning of day t and stores it in the warehouse first. Throughout day t, the warehouse delivers the part in bins to workstations upon receiving their requests. The line-side inventory status at workstation i (S_i^t) is determined by the line-side inventory of the previous day (S_i^{t-1}), the workstation demand (D_i^t), and the delivery amount from the warehouse.

Table 4.2 lists the order quantities for the four cases of information availability. Each variable or parameter in the table is in units of bins and $\lceil \cdot \rceil$ is the ceiling function.

First, in Case 1 the plant calculates the order quantity as the next-day demand (aggregated across all workstations) plus the difference between the minimum inventory and the existing inventories in the warehouse and at workstations. After that, the plant rounds up this amount to the closest bin quantity since the supplier can only ship the part to the plant in bins. By using this order quantity, the plant ensures that the remaining inventory (in the warehouse and at workstations) at the end of a day to be no less than the minimum inventory and no more than the minimum inventory plus one bin. We can interpret the minimum inventory as the inventory we have to cover the effect of bin delivery. Note that if we set the minimum inventory to be nonnegative, we have excessive inventory in the plant at the end of any given day. However, it does not guarantee that we would satisfy all workstation requests. As the example we describe in Figure 4.1, while a workstation is starving, the excessive inventory could be stored at other workstations due to the bin delivery restriction. In Case 2, we assume that we do not know the existing line-side inventory and, in turn, do not consider it in the order quantity calculation. The minimum inventory here has a different interpretation from Case 1. It represents the extra inventory, in addition to the existing line-side inventory, we have to cover the effect of bin delivery.

Cases 3 and 4 are the extensions of Cases 1 and 2, respectively, where we assume we do not know the exact amount of next-day demand but only the distribution. In Case 3, we have no less than the minimum inventory and no more than the minimum inventory plus one whole bin in the plant at the beginning of each day. Roughly speaking, the minimum inventory is the inventory we have to satisfy the

daily demand of all workstations and to cover the effect of bin delivery. On the other hand, in Case 4 we have at least the minimum inventory, but no more than the minimum inventory plus one whole bin, in the warehouse at the beginning of each day. Hence, the minimum inventory plus the existing line-side inventory is the inventory we have for daily demand and the bin delivery effect.

Table 4.2: Order quantities calculation of Cases 1 to 4. For example, at the end of day $t-1$, we calculate the order quantity (O^t) for Case 1 as the next-day demand (D^t) plus the minimum inventory (Y) minus the remaining warehouse inventory (Z^{t-1}) and line-side inventory (S^{t-1}), and round up to the closet bin quantity. Each variable is in units of bins.

O^t	Known line-side inventory	Unknown line-side inventory
Known next-day demand	Case 1 $\lceil D^t + Y - (Z^{t-1} + S^{t-1}) \rceil$	Case 2 $\lceil D^t + Y - Z^{t-1} \rceil$
Unknown next-day demand	Case 3 $\lceil Y - (Z^{t-1} + S^{t-1}) \rceil$	Case 4 $\lceil Y - Z^{t-1} \rceil$

Warehouse Inventory Expression

Given O^t, we can express the warehouse inventory (Z^t) in terms of Y, D^t, S^{t-1}, and S^t. For inventory conservation, the following equation must hold:

$$O^t + Z^{t-1} + S^{t-1} = D^t + Z^t + S^t. \tag{4.1}$$

It implies that supply plus existing inventory is equal to demand plus the resulting inventory. Let h be the amount resulting from the ceiling function in the order quantity calculation. For Case 1, we can rewrite the inventory conservation equation as:

$$D^t + Y - Z^{t-1} - S^{t-1} + h + Z^{t-1} + S^{t-1} = D^t + Z^t + S^t.$$

After some rearrangement, we obtain the next-day warehouse inventory:

$$Z^t = Y - S^t + h = \left\lceil Y - S^t \right\rceil.$$

We can express Z^t in terms of Y, D^t, S^{t-1}, and S^t for the other three cases using a similar derivation. Table 4.3 lists the expressions of Z^t for all four cases.

Table 4.3: Next-day warehouse inventory of Cases 1 to 4. Based on the order quantity and equation (4.1), we can express the next-day warehouse inventory (Z^t) in terms of the minimum inventory (Y), next-day demand (D^t), existing line-side inventory (S^{t-1}), and resulting line-side inventory (S^t). For example, in Case 1 Z^t is the ceiling of Y minus S^t. Each variable is in units of bins.

Z^t	Known line-side inventory	Unknown line-side inventory
Known next-day demand	Case 1 $\lceil Y - S^t \rceil$	Case 2 $\lceil Y + S^{t-1} - S^t \rceil$
Unknown next-day demand	Case 3 $\lceil Y - D^t - S^t \rceil$	Case 4 $\lceil Y - D^t + S^{t-1} - S^t \rceil$

4.3.2 Properties of a Single Workstation with Sufficient Supply

Here, we show that for a single workstation, say workstation i, with sufficient supply, there are some intriguing properties, including the uniform stationary distribution of the line-side inventory, the weak dependence of the line-side inventories of two consecutive days, and the independence of the workstation demand and the resulting line-side inventory. Sufficient supply means that we replenish the workstation whenever there is a need and the remaining inventory at the workstation at the end of a day is strictly less than one whole bin. We then use these properties to derive a short-cut formula for each of the four cases to estimate the minimum inventory needed for a given warehouse part availability target.

4.3.2.1 Uniformly Distributed Line-Side Inventory

In order to obtain the stationary distribution of the line-side inventory at a workstation with sufficient supply, we consider the workstation demand of day t

in units of pieces (denoted by d_i^t) and assume it can be represented by a discrete nonnegative random variable with probability mass function (p.m.f.) $f_{d_i^t}$ and sample space Ω. Likewise, we use lowercase notation s_i^t to denote the line-side inventory at the end of day t in units of pieces. Note that there is one-to-one mapping between D_i^t and d_i^t via a multiplier b, i.e., $d_i^t = b \cdot D_i^t$. There is a similar relation between S_i^t and s_i^t. The sufficient supply assumption implies d_i^t and s_i^t satisfy:

$$s_i^t = s_i^{t-1} + \left\lceil \frac{(d_i^t - s_i^{t-1})}{b} \right\rceil \cdot b - d_i^t. \tag{4.2}$$

Dividing equation (4.2) by b on the both sides, we obtain:

$$S_i^t = S_i^{t-1} + \left\lceil D_i^t - S_i^{t-1} \right\rceil - D_i^t. \tag{4.3}$$

Next, we show that in the case of a single workstation i with sufficient supply the support of s_i^t, which is a random variable since it is a function of d_i^t and s_i^{t-1}, is $\{0, 1, 2, \ldots, b-1\}$.

Lemma 4.3.1. *Assume a single workstation i has line-side inventory s_i^t pieces at the end of day t and workstation demand d_i^t pieces on day t. If there is sufficient supply for the workstation, then the support of s_i^t is $\{0, 1, 2, \ldots, b-1\}$, where b is the bin size. Furthermore, the support of S_i^t $(= s_i^t/b)$ is $\left\{0, \frac{1}{b}, \frac{2}{b}, \cdots, \frac{b-1}{b}\right\}$.*

Proof. We know that the following relation holds for workstation i with sufficient supply:

$$s_i^t = s_i^{t-1} + \left\lceil \frac{(d_i^t - s_i^{t-1})}{b} \right\rceil \cdot b - d_i^t = \left\lceil \frac{(d_i^t - s_i^{t-1})}{b} \right\rceil \cdot b - (d_i^t - s_i^{t-1}).$$

Let $x = d_i^t - s_i^{t-1}$, which is an integer since d_i^t and s_i^{t-1} are integers. For a positive integer b and an integer x, we have [17]:

$$\begin{aligned}
\left\lceil \frac{x}{b} \right\rceil &\Leftrightarrow \frac{x}{b} \leq \left\lceil \frac{x}{b} \right\rceil < \frac{x}{b} + 1 \\
&\Leftrightarrow x \leq \left\lceil \frac{x}{b} \right\rceil \cdot b < x + b \\
&\Leftrightarrow 0 \leq \left\lceil \frac{x}{b} \right\rceil \cdot b - x < b \\
&\Leftrightarrow 0 \leq \left\lceil \frac{x}{b} \right\rceil \cdot b - x \leq b - 1.
\end{aligned}$$

The non-strict inequality ($\leq b-1$) in the last relation is due to the fact that $\left\lceil \frac{x}{b} \right\rceil \cdot b - x$ is an integer. Hence, we obtain:

$$0 \leq s_i^t \leq b - 1.$$

Consequently, the support of s_i^t is $\{0, 1, 2, \ldots, b-1\}$. Since S_i^t maps to s_i^t one-to-one, we also have that the support of S_i^t is $\left\{0, \frac{1}{b}, \cdots, \frac{b-1}{b}\right\}$. □

From equation (4.2) we know that the value of s_i^t only depends on s_i^{t-1} (the previous state) and d_i^t (a time-homogeneous random variable) so that we can create a discrete time Markov Chain (DTMC) [29] for the line-side inventory, i.e., $\{s_i^t, t \geq 0\}$. According to Lemma 4.3.1, the DTMC has a finite state space $\{0, 1, \ldots, b-1\}$. We argue that the DTMC is irreducible and aperiodic since for a general distribution of workstation demand, it is possible to have any integral amount of inventory (from 0 to $b-1$ pieces) remaining at the workstation at the end of a day, regardless of the status of the previous day. Theorem 4.3.2 gives the unique stationary distribution for this irreducible and aperiodic DTMC.

Theorem 4.3.2. *The DTMC $\{s_i^t, t \geq 0\}$ has a discrete uniform stationary distribution.*

Proof. Denote the transition probability matrix of the DTMC by P, which has dimension $b \times b$. Let $P = [p_{jk}]$ where $0 \leq j \leq b-1$ and $0 \leq k \leq b-1$. The interpretation of p_{jk} is that it is the probability that the line-side inventory of day t is k, given the line-side inventory of day $t-1$ is j. We can obtain the value of p_{jk} as follows:

$$p_{jk} = \sum_{\{u \in \Omega : (u+j) \bmod b = k\}} f_{d_i^t}(u).$$

Due to the nature of a DTMC and the replenishment procedure, we know (i) $p_{jk} \geq 0, \forall (j,k)$ and $\sum_k p_{jk} = 1, \forall j$; (ii) $p_{jk} = p_{j+1,k+1}, \forall j \leq b-2, k \leq b-2$; (iii) $p_{b-1,k} = p_{0,k+1}, \forall k \leq b-2$; (iv) $p_{j,b-1} = p_{j+1,0}, \forall j \leq b-2$; and (v) $p_{00} = p_{b-1,b-1}$. Regardless of the specific value of p_{jk}, we can express the summation of any column k in P as:

$$\begin{aligned}
\sum_{j=0}^{b-1} p_{jk} &= p_{0k} + p_{1k} + p_{2k} + \cdots + p_{k-1,k} + p_{kk} + p_{k+1,k} + \cdots + p_{b-2,k} + p_{b-1,k} \\
&= p_{0k} + p_{0,k-1} + p_{0,k-2} + \cdots + p_{01} + p_{00} + p_{0,b-1} + \cdots + p_{0,k+2} + p_{0,k+1} \\
&= \sum_{l=0}^{b-1} p_{0l} \\
&= 1.
\end{aligned}$$

Therefore, we conclude that P is a doubly stochastic matrix. From the properties of a irreducible and aperiodic DTMC with a doubly stochastic matrix [48], we know the DTMC has a uniform stationary distribution. That is, s_i^t has a discrete uniform distribution with a support of $\{0, 1, 2, \ldots, b-1\}$ in stationarity. □

Furthermore, we know S_i^t and s_i^t have one-to-one mapping relation and the bin size of a part is usually large, e.g., 1,000 pieces. Hence, we further replace S_i^t in stationarity with the standard continuous uniform distribution, i.e., $S_i^t \sim U(0,1)$ when deriving short-cut formulae.

4.3.2.2 Dependence of Line-Side Inventories of Two Consecutive Days

The line-side inventories of two consecutive days (S_i^{t-1} and S_i^t) are related to each other via the workstation demand (D_i^t) and the ceiling function as in equation (4.3). Nonetheless, as we show below, these random variables are independent of each other.

Proposition 4.3.1. *Under the assumptions of Theorem 4.3.2, if* $p_{jk} = 1/b, \forall(j,k)$, *then* S_i^{t-1} *is independent of* S_i^t *in stationarity.*

Proof. Assume the random vector (s_i^{t-1}, s_i^t) has p.m.f. $f_{s_i^{t-1},s_i^t}$. In stationarity, $f_{s_i^{t-1},s_i^t}$ is:

$$f_{s_i^{t-1},s_i^t}(s_i^{t-1} = j, s_i^t = k) = \frac{1}{b} \cdot p_{jk}.$$

If $p_{jk} = \frac{1}{b}$, then

$$f_{s_i^{t-1},s_i^t}(s_i^{t-1} = j, s_i^t = k) = \frac{1}{b^2} = f_{s_i^{t-1}}(s_i^{t-1} = j) \cdot f_{s_i^t}(s_i^t = k).$$

Therefore, s_i^{t-1} and s_i^t are independent of each other in stationarity, implying the independence of S_i^{t-1} and S_i^t in stationarity due to the one-to-one mapping relation of S_i^t and s_i^t. □

For a demand distribution with a flat p.m.f. and a wide support compared to the bin size, the value of p_{ij} is close, instead of equal, to $1/b, \forall(j,k)$. As a result, we have $f_{s_i^{t-1},s_i^t}(s_i^{t-1}, s_i^t) \approx f_{s_i^{t-1}}(s_i^{t-1}) \cdot f_{s_i^t}(s_i^t)$, which implies S_i^{t-1} and S_i^t are very weakly dependent of each other in stationarity. This is not a precise measurement and/or approximation, but it provides us a ground to ignore the dependence between S_i^{t-1} and S_i^t when developing short-cut formulae. We examine the effect of such an assumption in numerical examples later.

4.3.2.3 Independence of Demand and the Resulting Line-Side Inventory

The workstation demand (D_i^t) and the resulting line-side inventory (S_i^t) need to satisfy equation (4.3). Nonetheless, we show that in stationarity they are independent of each other regardless of the demand distribution.

Proposition 4.3.2. *Under the assumptions of Theorem 4.3.2, D_i^t is independent of S_i^t in stationarity.*

Proof. Assume the random vector $(d_i^t,\, s_i^t)$ has p.m.f. $f_{d_i^t, s_i^t}$, which we can express as:

$$\begin{aligned}
f_{d_i^t, s_i^t}(d_i^t = l, s_i^t = k) &= \mathbb{P}(s_i^t = k \mid d_i^t = l) \cdot \mathbb{P}(d_i^t = l) \\
&= \left[\sum_{k=0}^{b-1} \mathbb{P}(s_i^t = k \mid d_i^t = l, s_i^{t-1} = j) \cdot \mathbb{P}(s_i^{t-1} = j)\right] \cdot \mathbb{P}(d_i^t = l) \\
&= \left[\sum_{j=0}^{b-1} \mathbb{P}(s_i^t = k \mid d_i^t = l, s^{t-1} = j)\right] \cdot \frac{1}{b} \cdot \mathbb{P}(d_i^t = l) \\
&= \frac{1}{b} \cdot f_{d_i^t}(d_i^t = l) \\
&= f_{d_i^t}(d_i^t = l) \cdot f_{s_i^t}(s_i^t = k).
\end{aligned}$$

We establish the second equality by using the law of total probability. The third equality is implied by Theorem 4.3.2 since in stationarity s_i^{t-1} has a discrete uniform distribution. In the bracket of the third equality, the value of the summation is unity because once s_i^{t-1} and d_i^t are known, s_i^t is determined. The summands in this expression are either zero or one with the only element equal to one is when $k = (j + l) \mod b$.

Hence, d_i^t and s_i^t are independent of each other in stationarity, implying the independence of D_i^t and S_i^t in stationarity due to the one-to-one mappings of D_i^t to d_i^t and S_i^t to s_i^t. □

4.3.3 Cases 1 and 2 with a Warehouse Part Availability of 1

For Cases 1 and 2, we assume we know the next-day workstation demand before ordering and use it in the order quantity calculation. It is reasonable to ask if there exists a minimum inventory that ensures the warehouse part availability equal to 1, i.e., the warehouse satisfies all workstation requests. We use Lemma 4.3.1 to derive the minimum inventory needed for ensuring a warehouse part availability of 1 for Cases 1 and 2.

Proposition 4.3.3. *Assume there are n workstations in Case 1 with the order quantity calculated as:*

$$O^t = \left\lceil D^t + Y - \left(Z^t + S^{t-1}\right) \right\rceil.$$

If $Y = n - 1$, then the warehouse part availability is 1.

Proof. From Lemma 4.3.1, we know if a workstation has sufficient supply, the resulting line-side inventory (S_i^t) is nonnegative and strictly less than a whole bin. Extending the result to n workstations with sufficient supply, we have:

$$\sum_{i \in I} S_i^t = S^t < n.$$

As we describe before, we can express the warehouse inventory of Case 1 as $Z^t = \lceil Y - S^t \rceil$. Letting $Y = n - 1$ and combining with the above strict inequality, we obtain:

$$\begin{aligned} -1 &< (n-1) - S^t \\ \Rightarrow 0 &\leq \left\lceil (n-1) - S^t \right\rceil = Z^t. \end{aligned}$$

Hence, we have $Z^t \geq 0, \forall t$, which implies the warehouse part availability is 1. □

Proposition 4.3.4. *Assume there are n workstations in Case 2 with the order quantity calculated as:*

$$O^t = \left\lceil D^t + Y - Z^t \right\rceil.$$

If $Y = n - 1$, then the warehouse part availability is 1.

Proof. From Lemma 4.3.1, we know a workstation with sufficient supply has a resulting line-side inventory that is nonnegative and strictly less than a whole bin. Extending the result to n workstations with sufficient supply, we have:

$$0 \leq \sum_{i \in I} S_i^t = S^t < n.$$

We can express the warehouse inventory of Case 2 as $Z^t = \lceil Y + S^{t-1} - S^t \rceil$. The above inequalities hold for any day t. Thus, we have the following inequality for $S^{t-1} - S^t$:

$$-n < S^{t-1} - S^t.$$

Letting $Y = n - 1$ and combining with the above strict inequality, we obtain:

$$\begin{aligned} -1 &< (n-1) + S^{t-1} - S^t \\ \Rightarrow 0 &\leq \left\lceil (n-1) + S^{t-1} - S^t \right\rceil = Z^t. \end{aligned}$$

Hence, we have $Z^t \geq 0, \forall t$, which implies the warehouse part availability is 1. □

4.3.4 Formulae for Minimum Inventory Needed

We use Theorem 4.3.2 and Propositions 4.3.1-4.3.4 given above to derive a short-cut formula for each of the four cases, estimating the minimum inventory needed for a certain level of warehouse part availability.

Case 1

Motivated by Theorem 4.3.2, we replace the line-side inventory of a workstation in stationarity with a standard continuous uniform random variable, i.e., $S_i^t \sim U_i(0,1)$. Then, we have Z^t for Case 1 as follows:

$$Z^t = \left\lceil Y - S^t \right\rceil \approx \left\lceil Y - \sum_{i \in I} U_i(0,1) \right\rceil \approx \left\lceil N\left(Y - \frac{n}{2}, \frac{n}{12}\right) \right\rceil.$$

The second approximation comes from replacing the summation of n independent and identically distributed (i.i.d.) standard continuous uniform random variables with a normal random variable [11]. The warehouse part availability is the same as the probability that Z^t is non-negative, or in other words, the probability that the normal random variable is greater than -1. Hence, if the target warehouse part availability is α, we can invert the normal distribution to obtain the corresponding Y as follows:

$$Y(\alpha) = \frac{n}{2} - 1 + \Phi^{-1}(\alpha)\left(\sqrt{\frac{n}{12}}\right).$$

However, if α is 1, the Y goes to infinity which contradicts Proposition 4.3.3. The reason is that we use a normal distribution, whose support is infinite, to approximate the distribution of the summation of n i.i.d. standard uniform random variables, whose support is finite. Thus, we modify the Y needed to:

$$Y(\alpha) = \min\left\{n - 1, \frac{n}{2} - 1 + \Phi^{-1}(\alpha)\left(\sqrt{\frac{n}{12}}\right)\right\}.$$

Case 2

Similar to Case 1, we replace S_i^{t-1} and S_i^t with two standard continuous uniform random variables. Then, we have Z^t for Case 2 as follows:

$$Z^t = \left\lceil Y + S^{t-1} - S^t \right\rceil \approx \left\lceil Y + \sum_{j \in I} U_j(0,1) - \sum_{i \in I} U_i(0,1) \right\rceil \approx \left\lceil N\left(Y, \frac{n}{6}\right) \right\rceil.$$

In the second approximation, we assume the independence of S^{t-1} and S^t, which is true only in the case of a single workstation with sufficient supply in stationarity and a certain demand distribution as we show in Proposition 4.3.1. However, if the demand distribution of a workstation has a flat p.m.f. and a wide support compared to the bin size, then S_i^{t-1} and S_i^t are very weakly dependent. As we mention before, we ignore the dependence to simplify the formula derivation and later use numerical examples to test the performance. Similar to Case 1, given the target warehouse part availability is α, we can obtain the corresponding Y by inverting the normal distribution as follows:

$$Y(\alpha) = -1 + \Phi^{-1}(\alpha)\left(\sqrt{\frac{n}{6}}\right).$$

Likewise, if α is 1, the Y goes to infinity for the same reason we describe in Case 1. According to Proposition 4.3.4, we modify the Y needed to:

$$Y(\alpha) = \min\left\{n-1, -1 + \Phi^{-1}(\alpha)\left(\sqrt{\frac{n}{6}}\right)\right\}.$$

Case 3

Instead of the exact amount, in Case 3 we assume we model the demand at workstation i by a normal distribution with mean μ_i and variance σ_i^2, i.e., $D_i^t \sim N(\mu_i, {\sigma_i}^2)$. We also assume the demand independence among workstations and that $\{D_i^1, D_i^2, \cdots, D_i^t\}, \forall i$, are i.i.d. sequences. Similar to Case 1, we replace S_i^t with a standard continuous uniform random variable. Then, we have Z^t for Case 3 as follows:

$$\begin{aligned} Z^t &= \left\lceil Y - D^t - S^t \right\rceil \\ &\approx \left\lceil Y - \sum_{i\in I} N(\mu_i, {\sigma_i}^2) - \sum_{i\in I} U_i(0,1) \right\rceil \approx \left\lceil N\left(Y - \mu - \frac{n}{2}, \sigma^2 + \frac{n}{12}\right)\right\rceil. \end{aligned}$$

The second approximation comes from substituting the summation of n i.i.d. standard continuous uniform random variables with a normal random variable and adding it to another independent normal random variable [11]. From Proposition 4.3.2, we know D^t and S^t are independent of each other in the case of a single workstation with sufficient supply in stationarity, which justifies ignoring the dependence of D^t and S^t in the above approximation. Likewise, given the warehouse part availability is α, the corresponding Y is as follows:

$$Y(\alpha) = \mu + \frac{n}{2} - 1 + \Phi^{-1}(\alpha)\left(\sqrt{\sigma^2 + \frac{n}{12}}\right).$$

Case 4

Like Case 3, by assuming normally distributed daily workstation demand and uniformly distributed line-side inventory, we have Z^t for Case 4 as follows:

$$\begin{aligned} Z^t &= \left\lceil Y - D^t + S^{t-1} - S^t \right\rceil \\ &\approx \left\lceil Y - \sum_{i \in I} N(\mu_i, {\sigma_i}^2) + \sum_{j \in I} U_j(0,1) - \sum_{i \in I} U_i(0,1) \right\rceil \approx \left\lceil N\left(Y - \mu, \sigma^2 + \frac{n}{6}\right) \right\rceil. \end{aligned}$$

We know D^t and S^{t-1} are independent of each other since $\{D_i^1, D_i^2, \cdots, D_i^t\}$ is an i.i.d. sequence. As we describe in Cases 2 and 3, we ignore the dependences of D^t, S^{t-1} and S^t to obtain the second approximation. Thus, if the target warehouse part availability is α, we have the corresponding Y as follows:

$$Y(\alpha) = \mu - 1 + \Phi^{-1}(\alpha)\left(\sqrt{\sigma^2 + \frac{n}{6}}\right).$$

We list the short-cut formulae of Y for the four cases in Table 4.4 for the readers' convenience.

We can see these derivations as an application of risk pooling. For Cases 1 and 2, the stochastic element is not workstation demand but line-side inventory.

Table 4.4: Short-cut formulae of Cases 1 to 4 for the minimum inventory needed for a α level warehouse part availability. For example, in Case 3, the minimum inventory (Y) needs to be at least the mean of the aggregated demand (μ) plus half of the number of workstations (n) minus 1 plus the value of inverting the standard normal distribution at α multiplied by the square root of the variance of the aggregated demand (σ^2) plus ($n/12$).

Y	Known line-side inventory	Unknown line-side inventory
Known next-day demand	Case 1 $\min\left\{n-1, \frac{n}{2}-1+\Phi^{-1}(\alpha)\left(\sqrt{\frac{n}{12}}\right)\right\}$	Case 2 $\min\left\{n-1, -1+\Phi^{-1}(\alpha)\left(\sqrt{\frac{n}{6}}\right)\right\}$
Unknown next-day demand	Case 3 $\mu+\frac{n}{2}-1+\Phi^{-1}(\alpha)\left(\sqrt{\sigma^2+\frac{n}{12}}\right)$	Case 4 $\mu-1+\Phi^{-1}(\alpha)\left(\sqrt{\sigma^2+\frac{n}{6}}\right)$

What is uncertain here is the location of inventory remaining at the end of a day, or in other words, the distribution of the remaining inventory (in the warehouse and at workstations). We pool the uncertainty of the line-side inventory from each workstation by summing n i.i.d. standard uniform random variables and then using the inverse of the standard normal distribution to obtain the inventory needed for a certain performance level. For Cases 3 and 4, in addition to stochastic line-side inventory, we also pool risk out of stochastic workstation demand. As a result, we add an additional μ to the formulae and σ^2 inside the square root function. Another thing worth noticing is that the minimum inventory needed for a target α grows with the number of workstations in a nonlinear manner as the result of risk pooling, i.e., the minimum inventory is not proportional to the number of workstations.

In addition to the dependence assumptions, we address other differences between the short-cut formulae and the stylized replenishment process model as follows. In the short-cut formulae, we assume that the line-side inventories (S_i^{t-1} and S_i^t) are

nonnegative at any workstation i, which is true when the workstation has sufficient supply all the time. However, the warehouse may not satisfy the request from the workstation if it runs out inventory. So S_i^{t-1} and/or S_i^t can be negative, which means the request is backordered.

Furthermore, we assume the warehouse inventory (Z^t) can be negative and the warehouse part availability is equal to the probability that Z^t is strictly less than 0. However, the warehouse inventory is nonnegative all the time in the stylized model since the warehouse cannot deliver a bin to a workstation if it has no inventory. Instead of putting backorders at workstations as in the stylized model, we think of this as putting them in the warehouse when deriving the short-cut formulae. It is not surprising that there exists a gap between the target warehouse part availability and the true availability in the stylized replenishment process. In the following section, we examine this gap by simulating the stylized replenishment process with Y calculated from the short-cut formulae as input.

4.4 Numerical Examples

We simulate the stylized replenishment process to test the performance of the short-cut formulae of Cases 1 to 4. For each case, we run tests for the number of workstations n being 15 and 30. In our simulation, each workstation has a normal distribution-like daily demand. To be more specific, we first generate normal random variates. The variate is rounded up to be workstation demand (in units of pieces) since the demand can not be a fraction. Also, if the rounded-up variate is negative, we then take the demand of the day to be zero. The means of the workstation demands are randomly determined by a discrete uniform distribution with a support

of $\{500, \ldots, 5000\}$. The standard deviation is set to be one quarter of the mean, i.e., $\sigma_i = \mu_i/4$, and the bin size is set to be 1,000 pieces. We simulate the stylized replenishment process for 2,600 days, which is about 10 years of business days. In addition to these base examples, we also perform sensitivity analysis on demand variation ($\sigma_i = \mu_i/3$ and $\sigma_i = \mu_i/5$) and bin size (500 and 2,000 pieces) to see the effects on the performance of the short-cut formulae.

4.4.1 Warm-Up Time Determination

We use the following two tests to determine a proper warm-up time period, which is requisite since we set all line-side inventories to be zero initially in every simulation.

First, we test the time needed for a single workstation with sufficient supply to have a transient distribution close to the uniform stationary distribution, given the initial line-side inventory is zero. We construct the transition matrix (P, as we describe in Theorem 4.3.2) of different combination of means (of the 30 workstations) and standard deviations ($\sigma_i = \mu_i/3$, $\mu_i/4$, and $\mu_i/5$) with different bin sizes (500, 1,000, and 2,000) and set the initial distribution of s_i^0 to be $\{1, 0, 0, \ldots, 0\}$. During the testing, we observe that the time needed for a workstation to have stationarity-like transient distribution depends on the bin size and the standard deviation of the workstation demand. Typically, the bigger the bin size compared to the mean demand and/or the smaller variation of the demand, the longer it takes for s_i^t to have a discrete uniform-like distribution.

After 52 days, the distribution of s_i^{52} is $\{\frac{1}{b}, \frac{1}{b}, \frac{1}{b}, \cdots, \frac{1}{b}\}$ at the fifth decimal point for all tested combinations. We acknowledge that in the simulation, not all

workstation requests will be fulfilled, i.e., some of the workstations do not have sufficient supply. So we actually use 260 days as a very conservative warm-up period.

Second, given $\alpha = 1$ in Cases 1 and 2 and 0.9999 in Cases 3 and 4, we simulate the system for 260 days for all four cases with various combinations of demand variations ($\sigma_i = \mu_i/3$, $\mu_i/4$, and $\mu_i/5$) and bin sizes (500, 1,000, and 2,000) to obtain the correlation coefficient of D^t and S^t. According to Proposition 4.3.2, D^t and S^t have zero correlation in stationarity. With 5,000 replications, all tested combinations pass the Student's t-test, indicating non-rejection of the null hypothesis that D^t and S^t have zero correlation [28]. Hence, we determine the warm-up period to be 260 days.

4.4.2 Results

4.4.2.1 Base Examples

We test the performance of the short-cut formulae for different α ranging from 0.775 to 1 (0.9999 for Cases 3 and 4) in increments of 0.025 by simulating the system for 2,600 days using 50 replications. The average of 50 simulated warehouse part availability estimates, denoted by $\bar{\alpha}$, serves as a performance metric. In addition, we define the upper bound relative gap as:

$$Gap_{UB}(\alpha) = \frac{UB - \alpha}{\alpha} \cdot 100\%,$$

and the lower bound relative gap as:

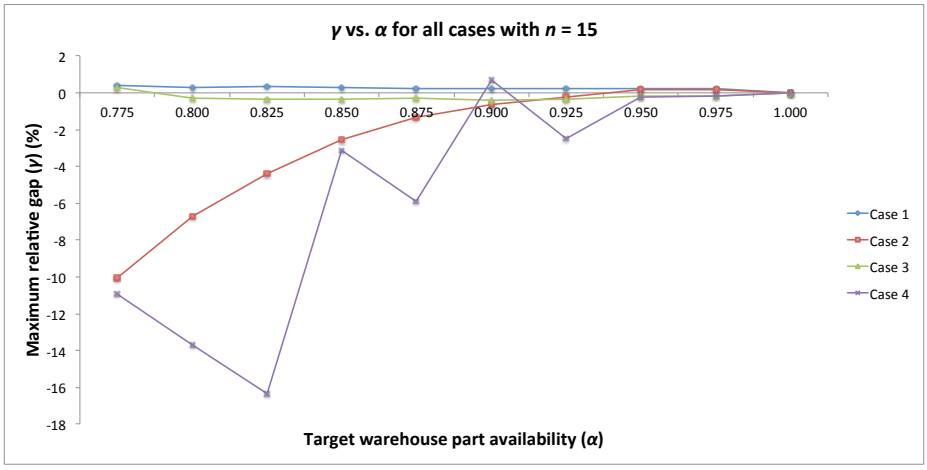
$$Gap_{LB}(\alpha) = \frac{\alpha - LB}{\alpha} \cdot 100\%,$$

where UB and LB denote the upper and lower bounds of the 95% confidence interval, respectively. As the maximum relative gap, denoted by γ, we pick one of Gap_{UB} and

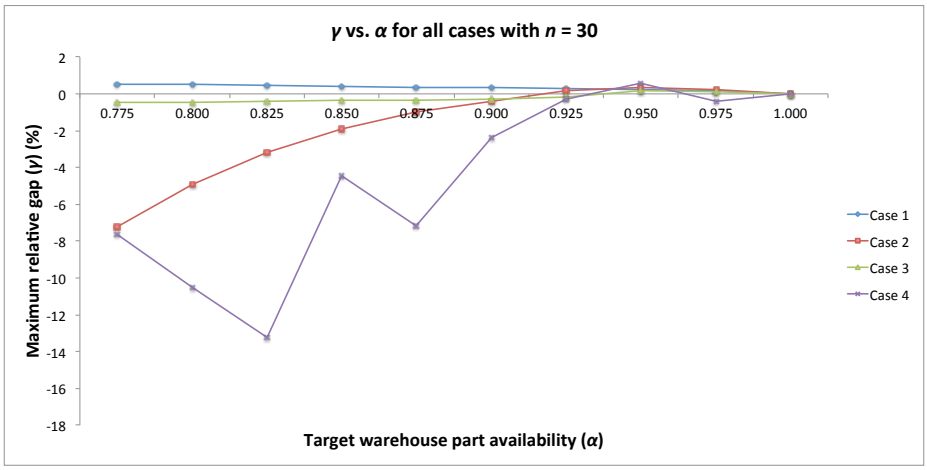
Gap_{LB} whose absolute value is larger. This γ value serves as another performance metric measuring the gap between α and $\bar{\alpha}$.

Figure 4.4 plots α versus γ for $n = 15$ in part (a) and $n = 30$ in part (b). We observe a tendency that the absolute value of γ gets smaller as α goes to 1 for all four cases with both $n = 15$ and 30, which implies that the short-cut formulae perform better as α is closer to 1. This is anticipated since the closer α is to 1, the more time S_i^t is nonnegative. We take a closer look at Cases 1 and 3 and see that for both $n = 15$ and 30, the absolute values of γ are close to 0 for any α tested, illustrating that the short-cut formulae of these two cases nicely predict the Y needed for a target α. On the other hand, in Cases 2 and 4 the absolute values of γ are much larger, especially when α is small. The largest absolute value of γ happens in Case 4. This is reasonable since we have the least information for the order quantity calculation and the loosest approximation (the independences of D^t, S^{t-1} and S^t) in deriving the short-cut formula for Case 4. In addition, we see that for Cases 2 and 4, most of the time γ is negative, which means the short-cut formulae tend to underestimate the Y needed for a given target α. Moreover, by comparing Cases 1 and 2, we can see that having the information of line-side inventory improves the performance of the short-cut formulae noticeably, which can also be observed from comparing Cases 3 and 4.

Table 4.5 lists the values of α, $\bar{\alpha}$, and γ for $n = 15$ in part (a) and $n = 30$ in part (b). We can see, both for $n = 15$ and 30, the absolute value of γ is less than 0.6% in Cases 1 and 3 where the line-side inventory is known, which means the true warehouse part availability of the system is 0.6% less than the target. Note that the short-cut formulae could under- or overestimate the Y needed for a target α since the



(a) $n = 15$



(b) $n = 30$

Figure 4.4: Maximum relative gap (γ) with different target warehouse part availability (α). We plot the results of 15 workstations in part (a) and that of 30 workstations in part (b). Each color represents the change of γ along with the growth of α for one case. A positive γ means the simulated average warehouse part availability is less than the target α, and vice versa. We can see that the short-cut formulae perform better when α is close to 1 for all four cases.

sign of γ alternates with the change of α. On the other hand, some of the absolute values of γ go beyond 10% in Cases 2 and 4 where the line-side inventory is unknown. Again, the sign of γ changes with different α. However, when α is small, e.g., 0.775 and 0.800, we have a large negative value of γ and a much smaller $\bar{\alpha}$ than α. It seems there exists a tendency in the short-cut formulae of Cases 2 and 4 to underestimate the Y needed for a small target α.

4.4.2.2 Sensitivity Analysis

In addition to the previous base examples, we perform sensitivity analysis on demand variation and bin size for all four cases.

Demand Variation

Figure 4.5 shows the results of how γ changes with α, given different demand variation, i.e., $\sigma_i = \mu_i/3$, $\mu_i/4$, and $\mu_i/5$, for $n = 15$ in part (a) and $n = 30$ in part (b). For Cases 1 and 3, there is no pattern that the short-cut formulae under- or overestimate the true warehouse part availability with the change of demand variation in both $n = 15$ and 30. Also, we can see almost no change in γ for Case 2 with both $n = 15$ and 30. For Case 4, we see more change when α is 0.775, especially with $n = 15$. The difference is smaller with $n = 30$. This can be explained by the fact that there is the most uncertainty (next-day demand plus line-side inventory) in Case 4 and the assumptions are far from reality when the warehouse part availability is small. Therefore, the short-cut formula has more errors in estimating the minimum inventory needed. The performance of the short-cut formulae improves when the number of workstations is larger since the normal approximations are more accurate for larger values of n. However, in general the absolute values of γ seem to be about

Table 4.5: Results of simulated average warehouse part availability ($\bar{\alpha}$) and the associated maximum relative gap (γ) for Cases 1 to 4, given different target warehouse part availability (α). We list the results of 15 workstations in part (a) and that of 30 workstations in part (b). Both in parts (a) and (b), the α is 1 in the last row in Cases 1 and 2 while it is 0.9999 in Cases 3 and 4. A negative γ implies $\bar{\alpha}$ is less than α. When α is 1, the $\bar{\alpha}$ is exactly 1 in Cases 1 and 2 since we use a sufficient minimum inventory to have a warehouse part availability of 1, as we describe in Proposition 4.3.4 and 4.3.3.

(a) $n = 15$

Target	Case 1		Case 2		Case 3		Case 4	
α	$\bar{\alpha}$	$\gamma(\%)$	$\bar{\alpha}$	$\gamma(\%)$	$\bar{\alpha}$	$\gamma(\%)$	$\bar{\alpha}$	$\gamma(\%)$
0.775	0.776	0.41	0.700	-10.07	0.775	0.27	0.694	-10.94
0.800	0.800	0.29	0.749	-6.69	0.800	-0.29	0.694	-13.72
0.825	0.826	0.37	0.791	-4.37	0.825	-0.32	0.694	-16.34
0.850	0.851	0.31	0.831	-2.55	0.849	-0.32	0.826	-3.14
0.875	0.875	0.24	0.865	-1.34	0.874	-0.32	0.826	-5.91
0.900	0.901	0.25	0.896	-0.64	0.898	-0.40	0.904	0.69
0.925	0.926	0.21	0.924	-0.25	0.923	-0.33	0.904	-2.49
0.950	0.951	0.26	0.950	0.16	0.949	-0.20	0.949	-0.26
0.975	0.976	0.21	0.976	0.20	0.974	-0.19	0.975	-0.16
1.000	1	0	1	0	1.000	0.01	1.000	0.01

(b) $n = 30$

Target	Case 1		Case 2		Case 3		Case 4	
α	$\bar{\alpha}$	$\gamma(\%)$	$\bar{\alpha}$	$\gamma(\%)$	$\bar{\alpha}$	$\gamma(\%)$	$\bar{\alpha}$	$\gamma(\%)$
0.775	0.777	0.52	0.721	-7.23	0.774	-0.48	0.719	-7.63
0.800	0.802	0.53	0.763	-4.94	0.799	-0.45	0.719	-10.52
0.825	0.827	0.49	0.801	-3.17	0.824	-0.42	0.719	-13.23
0.850	0.851	0.40	0.836	-1.89	0.849	-0.35	0.815	-4.44
0.875	0.876	0.36	0.868	-0.96	0.874	-0.33	0.815	-7.17
0.900	0.902	0.37	0.898	-0.42	0.899	-0.31	0.880	-2.40
0.925	0.926	0.27	0.925	0.19	0.925	-0.17	0.924	-0.29
0.950	0.951	0.20	0.952	0.33	0.950	0.15	0.954	0.56
0.975	0.976	0.16	0.977	0.25	0.975	0.11	0.972	-0.40
1.000	1	0	1	0	1.000	0.01	1.000	0.01

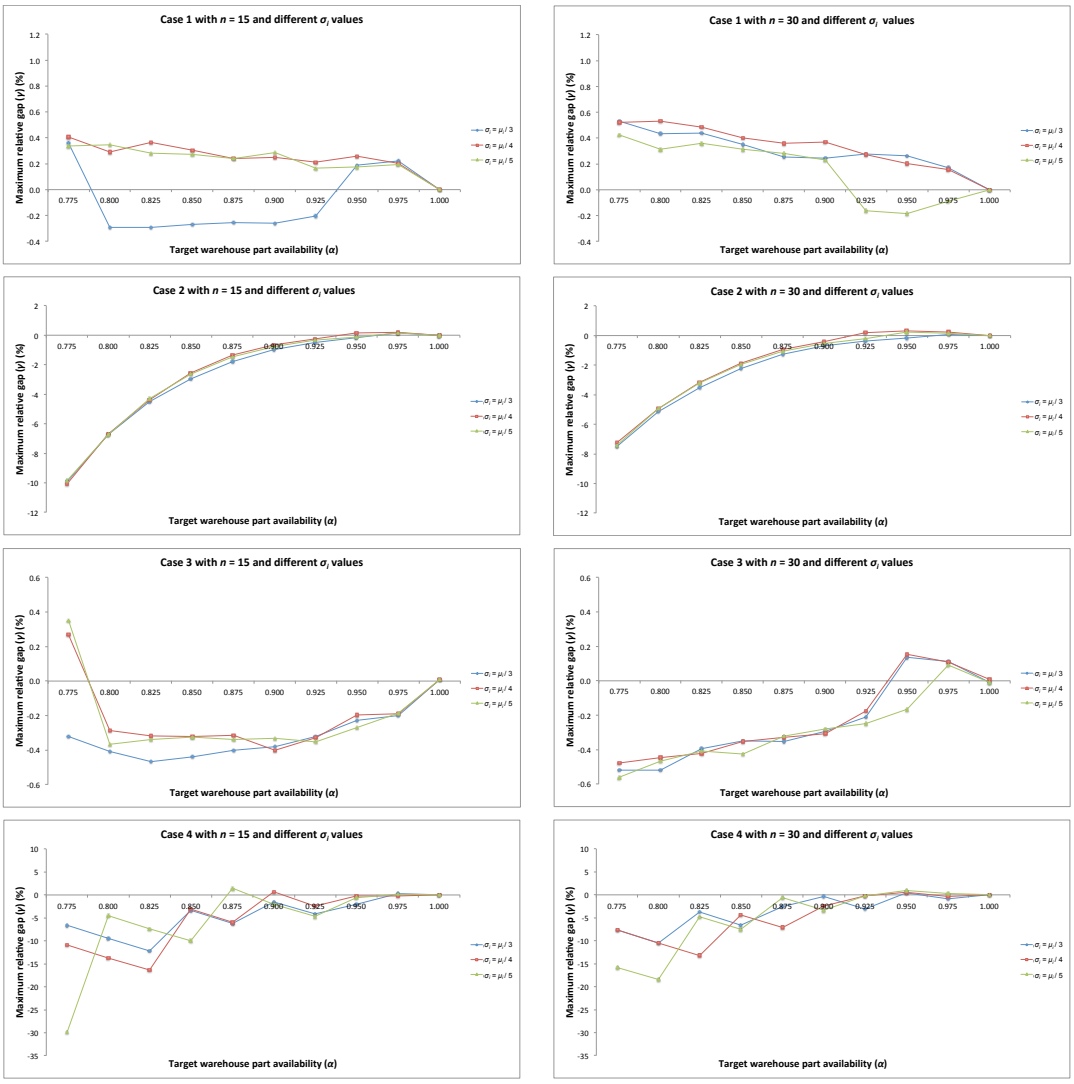
the same magnitude for all four cases, which implies the demand variation does not have significant effect on the performance of the short-cut formulae.

Bin Size

Figure 4.6 shows the results of how γ changes with α, given different bin size, i.e., $b = 500$, 1,000, and 2,000, again, for $n = 15$ in part (a) and $n = 30$ in part (b). Similar to the analysis of demand variation, for Cases 1 and 3 with both $n = 15$ and 30, the short-cut formulae may under- or overestimate the Y needed for a target α. We see no clear pattern on the change of γ with different b values. Although in $n = 30$ of Case 1 it seems the absolute value of γ becomes smaller with the growth of b for a fixed α, the absolute difference is less than 1%. On the other hand, for Cases 2 and 4, the value of b does not affect the value of γ, especially in Case 2, which has very little change for either $n = 15$ or 30. We observe some fluctuation of γ with different b values in Case 4, however, again, there is no clear pattern on what influence b has on γ. Nevertheless, we see that when α is small, the short-cut formulae tend to underestimate the Y needed as we mention before. In general, for all four cases we do not see noticeable effects of bin size on the change of the maximum relative gap, which implies that the bin size does not have notable influence on the performance of the short-cut formulae.

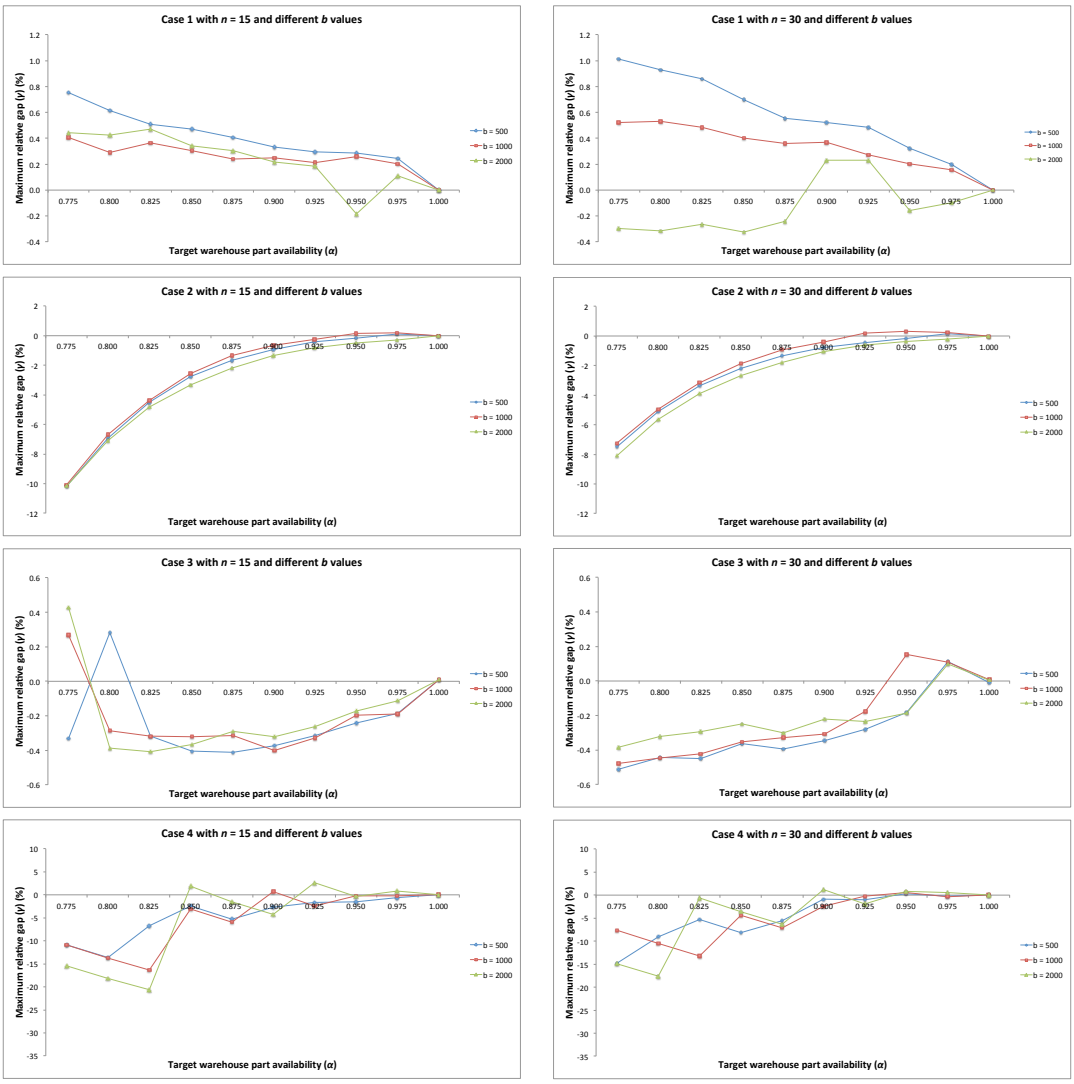
4.5 Discussion

From the base numerical examples, we see that the short-cut formulae predict the minimum inventory needed for a given target warehouse part availability better in Cases 1 and 3, compared to Cases 2 and 4. This implies that the information of line-side inventory is more valuable than that of exact next-day workstation demand.



(a) $n = 15$ (b) $n = 30$

Figure 4.5: Results of sensitivity analysis on demand variation. We list the results of 15 workstations for all four cases on the left-hand side of the figure and that of 30 workstations on the right-had side. A color represents one setting of standard deviation, e.g., blue represents the setting where the standard deviation of demand at a workstation (σ_i) is one third of its mean (μ_i). We see that the demand variation has very slight or no influence on the performance of the short-cut formulae for all four cases.



(a) $n = 15$ (b) $n = 30$

Figure 4.6: Results of sensitivity analysis on bin size. We list the results of 15 workstations for all four cases on the left-hand side of the figure and that of 30 workstations on the right-had side. A color represents one setting of bin size, e.g., blue represents the setting where the bin size is 500 pieces. We see that the bin size has very slight or no influence on the performance of the short-cut formulae for all four cases.

The reason is that in Cases 2 and 4, if the line-side inventory is negative, resulting from the inappropriate inventory distribution from bin delivery, we will not be able to catch it since we only check the warehouse inventory status. On the other hand, Cases 1 and 3 take the line-side inventory of the previous day into consideration when calculating the order quantity, so that the order quantity may be able to eliminate any existing backorders.

Furthermore, from the sensitivity analysis, we see demand variation and bin size have no significant influence on the performance of the short-cut formulae in all cases. It does not mean the demand variation and bin size have no influence at all. This analysis implies that for a reasonable bin size and demand variation, the formulae offer a good estimate for the minimum inventory needed. Surely, we can have some extreme examples in which the influence of bin size and demand variation is remarkable, e.g., bin size is equal to two pieces or demand variation is zero. However, such extreme examples seem to be pathological and are not the focus of this chapter.

In addition, the short-cut formulae have a maximum relative gap (at 95% confidence) less than 1% when the target warehouse part availability is greater than or equal to 0.950 for all of the base examples. In practice, requiring 95% satisfaction/service level is not uncommon and this implies the practicality of the short-cut formulae. Moreover, looking at the minimum inventory needed for a target availability, we observe the tradeoff between the warehouse part availability and the minimum inventory. Table 4.6 lists the Y needed for all of the base examples for $n = 15$ in part (a) and $n = 30$ in part (b). For Case 1, we need Y to be 14 bins for $n = 15$ and 29 for $n = 30$ to guarantee a warehouse part availability of 1, as we show in Proposition 4.3.3. However, if we are willing to sacrifice some availability, say 0.025,

then the Y decreases 38% for $n = 15$ and 41% for $n = 30$. This means that we can save inventory of 5.31 bins for $n = 15$ and 11.9 bins for $n = 30$. This decrease is even larger for Case 2, which is 85% for $n = 15$ and 88% for $n = 30$. For Cases 3 and 4 where next-day demand is unknown, we can observe the same significant decrease. The decrease in Y from $\alpha = 0.9999$ to 0.975, is about 5-6 bins for $n = 15$ and 7-8 bins for $n = 30$.

Note that the Y does not represent the average daily inventory in the plant, instead, it is part of the overall inventory and has different meanings in different cases. In reality, Case 2 has a slightly higher average daily inventory than Case 1 for the same target warehouse part availability since Case 2 has an additional source of uncertainty, i.e., line-side inventory. Likewise, Case 4 has a slightly higher daily inventory than Case 3 for the same reason. The magnitude of the decrease in Y from lowering the part availability from 1 down to 0.975 may seem small compared to the average daily inventory. Over time, significant savings could be realized since there may be hundreds of parts used within a plant, which is the case in the plant that we worked with.

In this chapter, we assume that the demand of a workstation is independent of other workstations. If the dependence among workstation demand can not be ignored, the short-cut formulae of Cases 1 and 2 should be still valid because the resulting line-side inventory is independent of demand in stationarity, as we show in Proposition 4.3.2, and the demand is explicitly considered when calculating the order quantity. For Cases 3 and 4, we may need to modify the value of σ^2 in the formulae to consider the demand dependence. On the other hand, if a workstation demand depends on the demand of the previous day significantly, then the short-cut formulae

Table 4.6: The minimum inventory (Y) needed from the short-cut formulae and the simulated average warehouse part availability ($\bar{\alpha}$) for Cases 1 to 4, given different target warehouse part availability (α). We list the results of 15 workstations in part (a) and that of 30 workstations in part (b). Both in parts (a) and (b), the α is 1 in the last row in Cases 1 and 2 while it is 0.9999 in Cases 3 and 4. We can see that the Y needed increases with α in a nonlinear manner.

(a) $n = 15$

Target	Case 1		Case 2		Case 3		Case 4	
α	$\bar{\alpha}$	Y (bins)	$\bar{\alpha}$	Y (bins)	$\bar{\alpha}$	Y (bins)	$\bar{\alpha}$	Y (bins)
0.775	0.776	7.34	0.700	0.19	0.775	50.71	0.694	43.35
0.800	0.800	7.44	0.749	0.33	0.800	50.98	0.694	43.64
0.825	0.826	7.54	0.791	0.48	0.825	51.28	0.694	43.96
0.850	0.851	7.66	0.831	0.64	0.849	51.61	0.826	44.31
0.875	0.875	7.79	0.865	0.82	0.874	51.98	0.826	44.69
0.900	0.901	7.93	0.896	1.03	0.898	52.40	0.904	45.14
0.925	0.926	8.11	0.924	1.28	0.923	52.90	0.904	45.68
0.950	0.951	8.34	0.950	1.60	0.949	53.56	0.949	46.38
0.975	0.976	8.69	0.976	2.10	0.974	54.58	0.975	47.45
1.000	1	14	1	14	1.000	60.23	1.000	53.43

(b) $n = 30$

Target	Case 1		Case 2		Case 3		Case 4	
α	$\bar{\alpha}$	Y (bins)	$\bar{\alpha}$	Y (bins)	$\bar{\alpha}$	Y (bins)	$\bar{\alpha}$	Y (bins)
0.775	0.777	15.19	0.721	0.69	0.774	95.95	0.719	81.16
0.800	0.802	15.33	0.763	0.88	0.799	96.31	0.719	81.55
0.825	0.827	15.48	0.801	1.09	0.824	96.71	0.719	81.98
0.850	0.851	15.64	0.836	1.32	0.849	97.15	0.815	82.44
0.875	0.876	15.82	0.868	1.57	0.874	97.63	0.815	82.96
0.900	0.902	16.03	0.898	1.87	0.899	98.20	0.880	83.56
0.925	0.926	16.28	0.925	2.22	0.925	98.87	0.924	84.28
0.950	0.951	16.60	0.952	2.68	0.950	99.75	0.954	85.21
0.975	0.976	17.10	0.977	3.38	0.975	101.10	0.972	86.65
1.000	1	29	1	29	1.000	108.62	1.000	94.67

may not perform well since the independence of workstation demand among days is a crucial assumption in Theorem 4.3.2.

The warehouse part availability is defined as number of days that the warehouse does not satisfy all workstation requests. Another interpretation of the metric is that it represents how frequently the plant needs to move the part between workstations manually. Like other fractional key performance indices, the warehouse part availability does not capture the number of unfulfilled requests and their magnitude. However, in the simulation results, including the base examples and sensitivity analysis, when the target warehouse part availability is close to 1, e.g., 0.950, the number of unfulfilled requests is small, compared to the number of total workstations, and the amounts of these requests are small. This means that even there are unfulfilled requests, minimal efforts should be required to maintain the production of the assembly lines.

Chapter 5

Conclusions

Stockpiling is one of the main ways of dealing with demand uncertainty and future demand surges. However, it comes with various costs, and hence stockpiling in an economic manner becomes an important topic receiving significant attention. Moreover, when demand exceeds supply, resource allocation becomes a major problem facing the decision maker. In this dissertation, we consider three problems of stockpiling and resource allocation: (i) stockpiling ventilators centrally and regionally for an influenza pandemic using risk pooling, (ii) allocating vaccines to priority groups at the geographic resolution of counties for an influenza pandemic seeking proportional fairness, and (iii) estimating extra inventory needed for class C parts due to demand aggregation and bin delivery. We summarize our solutions, major findings, and future work for each of these three topics as follows.

First, in Chapter 2 we optimize central and/or regional stockpiles of ventilators for an influenza pandemic to achieve different risk levels of unmet demand. We estimate the regional peak-week demands for ventilators based on a forecast of ILI hospitalizations using a dynamic linear model, the region-to-region correlation coefficient, the proportion of ILI patients requiring ICU care, the proportion of ICU patients requiring mechanical ventilation, and the proportion of ventilated patients requiring two weeks of ventilation. In addition to the regional demands for ventilators, our stockpile model also takes a wastage parameter as input, which accounts for

potential waste, or ineffectiveness, when distributing centrally held ventilators to regions. By choosing whether to fix the existing central or regional stockpiles as input, the stockpile model can help assess the performance of an existing central stockpile, the sufficiency of existing regional stockpiles, and the relative merits of central versus regional stockpiling.

By parameterizing a limit on the expected unmet demand in our stockpile model, we present the tradeoff between expected shortfall of ventilators and total stockpile, as well as the tradeoff between the probability of shortfall and total stockpile. We analyze a mild scenario based on data from the 2009 H1N1 pandemic in Texas. By scaling the mild scenario and using a larger proportion of hospitalizations requiring ICU care, we also examine moderate (like 1958/68) and severe (like 1918) scenarios. By fixing existing regional stockpiles in Texas, we find that no central stockpile is needed under the moderate scenario; however, there is a huge shortfall under the severe scenario. We also perform sensitivity analyses on the model's input parameters. Changing the ICU and/or ventilation proportions results in scaling our baseline stockpiling solutions. On the other hand, changing the two-week ventilation proportion is more subtle because consecutive weeks do not have identical hospitalizations. We provide a simplistic scaling rule to estimate the corresponding stockpiles using the baseline results. Moreover, the wastage proportion and region-to-region correlation coefficient, as well as the coefficient of variation of regional demands, affect the distribution of central and regional stockpiles. A lower value of the wastage parameter and a lower region-to-region correlation coefficient result in a larger central stockpile, as does a larger coefficient of variation.

In Chapter 2, we focus exclusively on ventilators because we have sufficient

data available for estimating regional levels of the corresponding supply and demand. However, the modeling framework can be readily extended to other critical resources, e.g., personal protective equipment and antivirals, should adequate data for these resources become available.

Second, in Chapter 3 we present an optimization-based framework for allocating available vaccine doses of different types to multiple priority groups at the geographic resolution of counties, maximizing proportionally fair coverage while keeping policy simplicity and regional equity in mind. In the system we consider, vaccine doses are assigned to two distribution systems, a pull-based system and a push-based system, to reach the public. Our first optimization model takes as input user-specified priority groups, weights for each county-priority group pair, suitability of different vaccine types for each priority group, pre-allocated vaccine doses from the pull-based system, and available vaccine doses reserved for the push-based system. Then, it provides the optimal coverage for each pair as output, seeking to bring all underserved county-priority group pairs to a proportionally fair level when weighted by their relative importance. The weights for each county-priority group pair reflect the user's desired relative coverage rates. Our second optimization model takes as input the optimal proportionally fair coverage rates from the first model and provides an optimal allocation for the push-based system according to two secondary objectives: policy simplicity and regional equity. Within the context of our first optimization model, we prove a formal result, establishing that the model's objective function, when minimized, ensures the desired notion of proportionally fair coverage under natural assumptions.

We take the vaccine distribution in Texas during the 2009 H1N1 pandemic as

a case study. At the time, four types of vaccines were distributed via three channels: RPs, LHDs, and HSRs. The first two channels are a pull-based distribution while the last one is a pushed-based distribution. We present results on how the 7% of total vaccine doses reserved by DSHS for discretionary allocation to HSR counties could bring the under-served county-priority group pairs within the 189 rural counties in Texas to a proportionally fair level of coverage after allocation of the RP and LHD doses. Among multiple optimal allocations, we select one with fewer priority group-vaccine type pairs for policy simplicity and with similar composition of vaccine types to health service regions for regional equity. We also perform sensitivity analysis on the portion of total doses reserved for HSR allocation. With a larger number of HSR doses, we can achieve a higher level of proportionally fair coverage in the 189 rural counties. Also, with a larger number of HSR doses we can effectively shrink the coverage gap between over-served county-priority group pairs and under-served ones. However, if the portion exceeds 7%, rural areas (the 189 counties eligible for HSR doses) may have more vaccine doses than the urban areas (the other 65 counties), which is an undesirable distribution.

The framework we describe in Chapter 3 focuses on vaccine doses because of the original motivation of equitable vaccine coverage across the 254 counties in the state of Texas. The analysis conducted in the chapter demonstrates the capability of an optimization-based framework using a one-time allocation. That said, it can be easily extended to a time-dynamic allocation in a rolling-horizon fashion. Furthermore, given a prediction of how influenza is spreading geographically over time, the framework can be used to allocate available vaccines. The framework might also extend to distribute other critical medical resources with complicated suitability of different types for each priority group, e.g., antivirals, where equitable coverage or

any desired relative coverage rate is the main objective.

Third, in Chapter 4 we build a stylized replenishment process model and an associated simulation model to investigate the effect of demand aggregation and bin delivery on class C part management for an engine assembly plant. Ordering small, relatively inexpensive parts from the suppliers and delivering them from a warehouse to workstations in bins saves material handling and delivery costs but causes an undesired side effect, i.e., uncertain line-side inventory distribution at workstations. This can result in not having the part at the right workstation when it is needed. According to information availability in the order quantity calculation, we consider four cases with two-dimensional uncertainty: next-day workstation demand and line-side inventory. By implementing a risk-pooling idea on the uncertain line-side inventory at workstations at the end of a day (and the uncertain next-day workstation demand if the demand is unknown), we derive valid short-cut formulae estimating the extra inventory (minimum inventory) needed for a given risk level of not satisfying all workstation requests due to aggregation of workstation demand and bin delivery.

Based on reasonable assumptions, we argue that the line-side inventory of a workstation with sufficient supply can be approximated as a standard continuous uniform random variable in stationarity when the bin size is large enough. Also, we show line-side inventories on two consecutive days are weakly dependent, and we show next-day workstation demand and the resulting line-side inventory are independent. Furthermore, we derive short-cut formulae to estimate the minimum inventory needed for a target risk level. The minimum inventory needed grows with the number of workstations in a nonlinear manner, i.e., the minimum inventory is not proportional to the number of workstations. This can be explained by the effect of risk pooling on

the uncertain line-side inventory at workstations. Based on numerical examples, we show the short-cut formulae can estimate the minimum inventory needed quite well, especially when the target warehouse part availability is above 0.95. Our sensitivity analysis shows that the variability of workstation demand and the size of bin do not have significant influence on the performance of the short-cut formulae.

Finally, in our models in which next-day demand is known when ordering, we expect that the correlation of demand among workstations has no effect on the short-cut formulae and the performance, because the workstation demand is well-informed in the order quantity calculation. On the other hand, in our models in which next-day demand is unknown when ordering, accounting for dependence among workstation demand requires slight modification to the short-cut formulae. However, if there is a noticeable correlation (or dependence) in workstation demand over time, the short-cut formulae may not be valid, and extending them to handle this situation would require further investigation. Another extension to this work that is worthy of further study is the analysis of two different ordering mechanisms: aggregated demand ordering with repacking before delivering the parts to workstations versus individual demand ordering without repacking, focusing on the tradeoff between the overall inventory level and the repacking cost. In terms of further related applications, the binned nature of high-demand class C parts may appear in some medical resources as well, e.g., multi-dose vial vaccines, and latex gloves and masks used as personal protective equipment. The results of Chapter 4 can serve as a starting point for further research considering the effect of package-quantity delivery on allocating critical medical resources during an influenza pandemic.

Bibliography

[1] American Association for Respiratory Care. Guidelines for Acquisition of Ventilators to Meet Demands for Pandemic Flu and Mass Casualty Incidents, 2008. `https://www.aarc.org/resources/vent_guidelines_08.pdf`, accessed June 6, 2014.

[2] American Association for Respiratory Care. The Strategic Nation Stockpile (SNS) Ventilator Training Program, 2014. `https://www.aarc.org/resources/sns_vent_training/`, accessed June 6, 2014.

[3] O. M. Araz, A. Galvani, and L. A. Meyers. Geographic Prioritization of Distributing Pandemic Influenza Vaccines. *Health Care Management Science*, 15(3):175–187, 2012.

[4] S. Axsäter. *Inventory Control*. Springer, New York, NY, USA, second edition, 2006.

[5] E. Baranoff, P. L. Brockett, and Y. Kahane. *Enterprise and Individual Risk Management*. Flat World Knowledge Inc., Washington, D. C., USA, 2009.

[6] D. Battini, M. Faccio, A. Persona, and F. Sgarbossa. Design of the Optimal Feeding Policy in an Assembly System. *International Journal of Production Economics*, 121(1):233–254, 2009.

[7] M. Baudin. *Lean Logistics: The Nuts and Bolts of Delivering Materials and Goods*. Productivity Press, New York, NY, USA, 2004.

[8] D. Bertsekas and R. Gallager. *Data Networks*. Prentice Hall, Englewood Cliffs, NJ, USA, second edition, 1992.

[9] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, CB2 8RU, UK, 2004.

[10] A. C. Caputo and P. M. Pelagagge. A Methodology for Selecting Assembly Systems Feeding Policy. *Industrial Management and Data Systems*, 111(1):84–112, 2011.

[11] G. Casella and R. L. Berger. *Statistical Inference*. Cengage Learning, Pacific Grove, CA, USA, second edition, 2002.

[12] A. Coluccia, A. D'Alconzo, and F. Ricciato. On the Optimality of Max-Min Fairness in Resource Allocation. *Annals of Telecommunications*, 67(1-2):15–26, 2012.

[13] L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, NY, USA, 1986.

[14] A. Ercole, B. L. Taylor, A. Rhodes, and D. K. Menon. Modeling the Impact of an Influenza A/H1N1 Pandemic on Critical Care Demand from Early Pathogenicity Data: the Case for Sentinel Reporting. *Anaesthesia*, 64(9):937–941, 2009.

[15] Ethics Subcommittee of the Advisory Committee to the Director, U.S. Centers for Disease Control and Prevention. Ethical Considerations for Decision Making Regarding Allocation of Mechanical Ventilators during a Severe Influenza Pandemic or Other Public Health Emergency, 2011. `http://www.cdc.gov/od/`

science/integrity/phethics/docs/Vent_Document_Final_Version.pdf, accessed May 16, 2014.

[16] J. Figueira, S. Greco, and M. Ehrgott. *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer-Verlag, New York, NY, USA, 2006.

[17] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley Professional, Reading, MA, USA, second edition, 1998.

[18] C. K. Grissom, S. M. Brown, K. G. Kuttler, J. P. Boltax, J. Jones, A. R. Jephson, and J. F. Orme, Jr. A Modified Sequential Organ Failure Assessment Score for Critical Care Triage. *Disaster Medicine and Public Health Preparedness*, 4(4):277–284, 2010.

[19] R. Hanson and A. Brolin. A Comparison of Kitting and Continuous Supply in In-Plant Materials Supply. *International Journal of Production Research*, 51(4):979–992, 2013.

[20] R. Hanson and C. Finnsgård. Impact of Unit Load Size on In-Plant Material Supply Efficiency. *International Journal of Production Economics*, 147(Part A):46–52, 2014.

[21] W. J. Hopp and M. L. Spearman. *Factory Physics*. McGraw-Hill, New York, NY, USA, third edition, 2008.

[22] S. Y. Hua and D. J. Johnson. Research Issues on Factors Influencing the Choice of Kitting versus Line Stocking. *International Journal of Production Research*, 48(3):779–800, 2010.

[23] H. Huang, O. M. Araz, D. P. Morton, and L. A. Meyers. Demand Forecasting for Medical Resources. Technical report, Texas Department of State Health Services, 2011.

[24] H. Huang, O. M. Araz, D. P. Morton, and L. A. Meyers. Stockpiling Ventilators for Pandemic Influenza. Technical report, Texas Department of State Health Services, 2011.

[25] H. Huang, B. Singh, G. P. Johnson, D. P. Morton, and L. A. Meyers. Decision-Support Tool for Allocating Pandemic Influenza Vaccines to Texas Health Service Regions and Counties. Technical report, Texas Department of State Health Services, 2013.

[26] M. J. Keeling and P. J. White. Targeting Vaccination against Novel Infections: Risk, Age and Spatial Structure for Pandemic Influenza in Great Britain. *Journal of the Royal Society Interface*, 8(58):661–670, 2010.

[27] F. Kelly. Charging and Rate Control for Elastic Traffic. *European Transactions on Telecommunications*, 8(1):33–37, 1997.

[28] M. G. Kendall and A. Stuart. *The Advanced Theory of Statistics, Volume 2: Inference and Relationship*. Charles Griffin and Co., London, England, UK, fourth edition, 1979.

[29] V. G. Kulkarni. *Modeling and Analysis of Stochastic Systems*. CRC Press, Boca Raton, FL, USA, second edition, 2010.

[30] J.-Y. Le Boudec. Rate Adaption, Congestion Control and Fairness: A Tutorial, 2012. `http://ica1www.epfl.ch/PS_files/LEB3132.pdf`, accessed on June 6,

2014.

[31] V. Limère, H. V. Langeghem, M. Goetschalckx, E.-H. Aghezzaf, and L. F. McGinnis. Optimising Part Feeding in the Automotive Assembly Industry: Deciding between Kitting and Line Stocking. *International Journal of Production Research*, 50(15):4046–4060, 2012.

[32] J. L. Logan. Disparities in Influenza Immunization among US Adults. *Journal of the National Medical Association*, 101(2):161–166, 2009.

[33] R. T. Marker and J. S. Arora. Survey of Multi-Objective Optimization Methods for Engineering. *Structural and Multidisciplinary Optimization*, 26(6):369–395, 2004.

[34] L. Matrajt, M. E. Halloran, and I. M. Longini, Jr. Optimal Vaccine Allocation for the Early Mitigation of Pandemic Influenza. *PLoS Computational Biology*, 9(3):e1002964, 2013.

[35] J. Medlock and A. P. Galvani. Optimizing Influenza Vaccine Distribution. *Science*, 325(5948):1705–1708, 2009.

[36] Minnesota Department of Health. Patient Care Strategies for Scarce Resource Situations, 2013. `http://www.health.state.mn.us/oep/healthcare/standards.pdf`, accessed May 16, 2014.

[37] J. Mo and J. Walrand. Fair End-to-End Window-Based Congestion Control. *IEEE/ACM Transactions on Networking*, 8(5):556–567, 2000.

[38] G. Neyman and C. B. Irvin. A Single Ventilator for Multiple Simulated Patients to Meet Disaster Surge. *Academic Emergency Medicine*, 13(11):1246–1249, 2006.

[39] Texas Department of State Health Services. H1N1 Vaccine Doses by County to Texas Registered Providers, 2010. http://www.dshs.state.tx.us/txflu/H1N1-Doses-Providers.pdf, accessed on May 16, 2014.

[40] Texas Department of State Health Services. H1N1 Vaccine Doses to Texas DSHS Regional Offices, 2010. http://www.dshs.state.tx.us/txflu/H1N1-Doses-Regions.pdf, accessed on May 16, 2014.

[41] Texas Department of State Health Services. H1N1 Vaccine Doses to Texas Local Health Departments, 2010. http://www.dshs.state.tx.us/txflu/H1N1-Doses-LHD.pdf, accessed on May 16, 2014.

[42] Texas Department of State Health Services. Internal Dashboard Report for H1N1. Technical report, Texas Department of State Health Services, 2010.

[43] World Health Organization. WHO Recommendations for the Post-Pandemic Period, 2010. http://www.who.int/csr/disease/swineflu/notes/briefing_20100810/en/, accessed on May 16, 2014.

[44] World Health Organization. *Report of the WHO Pandemic Influenza A(H1N1) Vaccine Deployment Initiative*. World Health Organization, Geneva, Switzerland, 2012.

[45] L. Paladino, M. Silverberg, J. G. Charchaflieh, J. K. Eason, B. J. Wright, N. Palamidessi, B. Arquilla, R. Sinert, and S. Manoach. Increasing Ventilator Surge Capacity in Disasters: Ventilation of Four Adult-Human-Sized Sheep on a Single Ventilator with a Modified Circuit. *Resuscitation*, 77(1):121–126, 2008.

[46] T. Powell, K. C. Christ, and G. S. Birkhead. Allocation of Ventilators in a Public Health Disaster. *Disaster Medicine and Public Health Preparedness*, 2(1):20–26, 2008.

[47] T. Rebmann and A. Zelicoff. Vaccination against Influenza: Role and Limitations in Pandemic Intervention Plans. *Expert Review of Vaccines*, 11(8):1009–1019, 2012.

[48] S. M. Ross. *Introduction to Probability Models*. Academic Press, Burlington, MA, USA, 10th edition, 2010.

[49] D. Simchi-Levi, P. Kaminsky, and E. Simchi-Levi. *Designing and Managing the Supply Chain: Concepts, Strategies, and Cases*. McGraw-Hill, New York, NY, USA, 2000.

[50] P. Smetanin, D. Stiff, A. Kumar, P. Kobak, R. Zarychanski, N. Simonsen, and F. Plummer. Potential Intensive Care Unit Ventilator Demand/Capacity Mismatch due to Novel Swine-origin H1N1 in Canada. *The Canadian Journal of Infectious Diseases and Medical Microbiology*, 20(4):115–123, 2009.

[51] M. J. Sobel. Risk Pooling. In *Building Intuition: Insights from Basic Operations Management Models and Principles*, pages 155–174. Springer-Verlag, New York, NY, USA, 2008.

[52] D. Stiff, A. Kumar, N. Kissoon, R. Fowler, P. Jouvet, P. Skippen, P. Smetanin, M. Kesselman, and S. Veroukis. Potential Pediatric Intensive Care Unit Demand/Capacity Mismatch due to Novel pH1N1 in Canada. *Pediatric Critical Care Medicine*, 12(2):e51–e57, 2011.

[53] C. Stroud, L. Nadig, B. M. Altevogt, and Rapporteurs. *The 2009 H1N1 Influenza Vaccination Campaign: Summary of a Workshop Series*. The National Academies Press, Washington, D. C., USA, 2010.

[54] J. Sutton and K. Tierney. Disaster Preparedness: Concepts, Guidance, and Research. In *Fritz Institute Assessing Disaster Preparedness Conference*, Sebastopol, CA, USA, November 2006.

[55] Texas Department of State Health Services. Texas Aggregate Surveillance Summary-Novel Influenza A H1N1, week ending 12/26/09, 2009. http://www.dshs.state.tx.us/txflu/TX-cumulative-age-archive.shtm, accessed on May 21, 2014.

[56] Texas Department of State Health Services. Texas Health Service Regions, 2011. http://www.dshs.state.tx.us/regions/state.shtm, accessed on May 16, 2014.

[57] Texas Department of State Health Services. TSA Bed-Count Dashboard Display, February-March 2011. Technical report, Texas Department of State Health Services, 2011.

[58] Texas Department of State Health Services. Pandemic Mitigation, 2013. https://www.preparingtexas.org/Resources/documents/2013%20Conference%20Presentations/Pandemic%20Mitigation.pdf, accessed May 16, 2014.

[59] Texas State Government. Texas Administrative Code, Title 25, Part 1, Chapter 157, Subchapter G, Rule §157.122, 2004. http://info.sos.state.tx.us/pls/pub/readtac$ext.TacPage?sl=R&app=9&p_dir=&p_rloc=&p_tloc=&p_ploc=&pg=1&p_tac=&ti=25&pt=1&ch=157&rl=122, accessed on May 21, 2014.

[60] J. W. Timbie, J. S. Ringel, D. S. Fox, F. Pillemer, D. A. Waxman, M. Moore, C. K. Hansen, A. R. Knebel, R. Ricciardi, and A. L. Kellermann. Systematic Review of Strategies to Manage and Allocate Scarce Resources during Mass Casualty Events. *Annals of Emergency Medicine*, 61(6):677–689, 2013.

[61] M. Uchida. Information Theoretic Aspects of Fairness Criteria in Network Resource Allocation Problems. In *Proceedings of the 2nd International Conference on Performance Evaluation Methodologies and Tools*, Nantes, France, October 2007.

[62] U.S. Centers for Disease Control and Preventioin. Strategic National Stockpile (SNS), 2012. `http://www.cdc.gov/phpr/stockpile/stockpile.htm`, accessed on June 16, 2014.

[63] U.S. Centers for Disease Control and Prevention. 2009 H1N1 Vaccination Recommendations, 2009. `http://www.cdc.gov/h1n1flu/vaccination/acip.htm`, accessed on May 16, 2014.

[64] U.S. Centers for Disease Control and Prevention. Seasonal Influenza Vaccine Supply and Distribution in the United States: Questions & Answers, 2013. `http://www.cdc.gov/flu/about/qa/vaxdistribution.htm`, accessed on May 16, 2014.

[65] U.S. Department of Health and Human Services. HHS Pandemic Influenza Plan, 2005. `http://www.flu.gov/planning-preparedness/federal/hhspandemicinfluenzaplan.pdf`, accessed on May 21, 2014.

[66] U.S. Department of Health and Human Services. Pandemic Flu History, 2014. `http://www.flu.gov/pandemic/history/index.html`, accessed on May 21, 2014.

[67] U.S. Department of Health and Human Services and U.S. Department of Homeland Security. Guidance on Allocation and Targeting Pandemic Influenza Vaccine, 2014. http://www.flu.gov/images/reports/pi_vaccine_allocation_guidance.pdf, accessed on May 16, 2014.

[68] U.S. Homeland Security Council. National Planning Scenarios, Version 21.3 Final Draft, 2006. https://publicintelligence.net/national-planning-scenarios-version-21-3-2006-final-draft/, accessed on May 21, 2014.

[69] J. Wilgis. Strategies for Providing Mechanical Ventilation in a Mass Casualty Incident: Distribution versus Stockpiling. *Respiratory Care*, 53(1):96–100, 2008.

[70] J. T. Wu, S. Riley, and G. M. Leung. Spatial Considerations for the Allocation of Pre-Pandemic Influenza Vaccination in the United States. *Proceedings of the Royal Society B: Biological Sciences*, 274(1627):2811–2817, 2007.

[71] X. Zhang, M. I. Meltzer, and P. M. Wortley. FluSurge 2.0: a Manual to Assist State and Local Public Health Officials and Hospital Administrators in Estimating the Impact of an Influenza Pandemic on Hospital Surge Capacity (Beta Test Version). Technical report, U.S. Centers for Disease Control and Prevention, 2005. http://www.cdc.gov/flu/pandemic-resources/tools/downloads/flusurge2.0_manual_060705.pdf, accessed on May 21, 2014.

[72] X. Zhang, M. I. Meltzer, and P. M. Wortley. FluSurge-A Tool to Estimate Demand for Hospital Services during the Next Pandemic Influenza. *Medical Decision Making*, 26(6):617–623, 2006.